Informatics and the Environmental Sciences

Karen S. Baker
Scripps Institution of Oceanography
University of California, San Diego

Oceanographic scientific sampling depends upon a national research vessel infrastructure; similarly, data require an information infrastructure. This photo is a reminder that ships are launched when their construction is complete. The challenge faced in working toward responsible data stewardship is that our information infrastructure – our understanding of data and concepts pertinent to informatics - is still in development.

Scripps Institution of Oceanography Technical Report
June 2005

Executive Summary

This report is on the topic of informatics and its relations to scientific research and *data* - rich, multi-faceted data that represent the earth and environmental systems. Data travel from field and laboratory into collections, repositories and archives. Just as data are a scientific resource, so too the work carried out with data and their organization is a resource for the environmental sciences.

Informatics is concerned with the stewardship of data, that is, with the tending of data and its flow, the design of information systems and their interfaces, and the growth of infrastructure given a distributed variety of data arenas. Enacted at the intersection of information science, environmental science and social science, informatics is evolving as we learn more about information environments and arrangements of human and technical systems. Five informatics 'good practices' are identified in this report:

Informatics Good Practices

1. Incorporate data problem formulation and data scoping early in the scientific planning process.
2. Recognize articulation, translation, negotiation and mediation as central to work with data.
3. Partner with appropriate information professionals for data work.
4. Create collaboration opportunities as well as coordination mechanisms for community work.
5. Recognize informatics as conducting research while carrying out information management.

The realm of informatics ranges across spatial, temporal, and organizational scales, weaving together diverse configurations, stretching over physical, digital and conceptual spaces. Many salient topics about data care remain to be discovered or investigated: data classification and provenance; data organization and modeling; data migration and data exchange; data assurance and quality control; data mediation and integration. Along with the development of roles for information professionals, we are learning about the dynamics of information environments, communities, and networks.

Informatics is happening. As we transition from use of 'my data' to 'our data', changes occur in data, collaborative, and scientific practices. Informatics provides new approaches and tools of interest to environmental scientists, information professionals, and social scientists alike.

I am an information manager privileged to work with several long-term, interdisciplinary projects within the Integrative Oceanography Division (IOD) at Scripps Institution of Oceanography and as an affiliate of the Science Studies Program at UCSD with its dynamic mix of communication, sociology, history and philosophy. Regarding my agenda with informatics, it is twofold: to be a responsible data steward and to partner with environmental science researchers by creating a contemporary information environment that supports concurrently the practice of information management and the inquiry of informatics research.

Karen S. Baker
June 06, 2005

Preface

Why read this report? Because if you are in science today, sometime recently you have likely experienced or witnessed a ‘data avalanche’ or experienced an ‘information overload’. You may have asked yourself, “What am I (are we) (will they be) doing with all this data?” Data old and new, data orderly and messy, data manually collected and data streaming so fast and from so many sources that “drowning in data” will come to mean not using the data and never rescuing undocumented ‘old’ data. Unless, perhaps, an approach is taken that enlists not solely the data collectors and technologists but, centrally and at the earliest development stages, takes into account the work of data managers, data technicians, information managers, data analysts, data scientists, information scientists, data users as well as administrators and the public. With this spectrum of participants engaged in dialogue and co-design, continuing assessment and ongoing refinement of designed entities is possible. Information systems can store and retrieve data according to local research needs while meeting growing long-term needs but we envision information systems as just one aspect of a comprehensive information environment approach. How might this approach be initiated? Well, read on…

This report presents an exploration of informatics within the context of contemporary environmental science environments with their attendant configurations of technology, local institutional settings, and broader community arrangements. Emerging from a voyage of discovery we are finding an operative base camp - a conceptual perspective - from which to develop effective approaches that provide for information management and infrastructure at a time when observational field science programs are transitioning to address multi-scale, interdisciplinary research questions collaboratively.

To insure that data remain available and useful, collections of data require some form of management. Because data are heterogeneous, this precludes a one-step or even a multi-step process. Data handling today entails an array of interdependent efforts that include collection and curation for both immediate use and future reuse of data. The goal of this report is to identify the work associated with establishing and maintaining data collections including the collaborative arrangements, the information infrastructure and the design processes required to support such collections. This work contributes to development of a vision for ocean sciences that we have called “Ocean Informatics”.

Data and information management activities are frequently a feature associated with data collecting although they are not always recognized as such. This report will set the stage for taking steps toward a long-term informatics approach to data stewardship within an organization as a cooperative undertaking that assumes information environments as learning environments are of central import to sustainable efforts.

Table of Contents

Executive Summary
Preface

1. The Context of Informatics

This report considers informatics and its relations to data at a time when our understanding of data and the work of organizing and managing them over the long-term is in transition and undergoing rapid development. Within this report working definitions of informatics and associated terms are drawn together, boxed as highlights, and placed for convenience in a report glossary (Appendix 7.1). The definitions are working definitions because informatics is an emergent field with a developing vocabulary.

1.1 A Brief History

The term “informatics” was born of the computer age but conceived from the words “inform” and “information” during the 14th and 15th centuries, respectively (OED 1981). The human desire to learn about and learn from observations and record events has probably been a salient characteristic of humankind way back into prehistoric times. Historically, joining together and delivering such ‘facts’ produced information and, once relationships had been perceived among the elements, a broader and deeper understanding would produce knowledge and expertise.

As modern science became coupled with the computational - the storage, retrieval and visualization potentials of all kinds of new tools, the volume and intricacy of data sets has burgeoned beyond most humans’ ability to imagine. By 1960 “information science” was being used to mean “the collection, classification, storage, retrieval, and dissemination of recorded knowledge treated both as a pure and as an applied science.”(Webster’s 1991)

About the same time, “informatique” was adopted by the French to connote the application of computing to the communication processes, i.e., dissemination, used by scientists to share information and data among themselves. France and other Western European countries viewed informatique as an application of computer science and subsequently a term to be used in place of computer science. Russians, however, used “informatika” to include a broader perspective. They saw this emerging field as a social science that would consider the use of technology within many distinct scientific communities and, of necessity, would also consider the interactions of technology with people and their organizational structures.

Today, in the United States, “informatics” is used in a variety of senses, often associated with information science, particularly with technology, data and communication. Informatics can incorporate interdependent associations resulting in scholarly literature on data work, information systems, data reuse, information management, digital preservation, information technology, and infrastructure[1] as well as societal interactions with all of the preceding elements. To observe the processes inherent in all these fields, sociotechnical research, social informatics, and infrastructure studies[2] draw upon fields such as Science Studies (SS), Science, Technology & Society (STS), Human Computer Interface (HCI), Participatory Design (PD), Infrastructure Studies (IS), and Information, Communication & Technology (ICT).

As contemporary scientific sampling and inquiry increase in scope, conceptual frameworks for data and informatics must also broaden. Interdisciplinary teams are one strategy for working at broad and/or large scales requiring input from multiple disciplines. Collaboration across disciplines is requisite for framing as well as addressing many contemporary informatics issues. Inevitably such teaming brings with it basic as well as nuanced issues of organizational refocus, community differences, and communication challenges. For example, the American medical community has used "medical informatics" to refer to management of medical data while the British use the term to connote "the place where health, information and computer sciences, psychology, epidemiology and engineering intersect." (BMIS).

Among the life sciences, "bioinformatics" began as the branch of informatics dealing with the massive data handling demands of the Human Genome Project. "Medical bioinformatics" came to incorporate the further knowledge that could be drawn from the genetic data coupled with clinical experience. "Ecoinformatics" emerged as informatics began to be used within ecological and environmental sciences. Contemporary uses of "informatics" conjoined with a specific domain tend to draw in domain specific elements of understandings about knowledge, knowledge-making and interface with technology in addition to data and information management.

Within the business community, the focus moves from data to information where information management is defined as "the application of management principles to the acquisition, organization, control, dissemination and use of information relevant to the effective operation of organizations of all kinds. 'Information' here refers to all types of information of value. Information management deals with the value, quality, ownership, use and security of information in the context of organizational performance." (Wilson, 1997)

Across the environmental sciences, cross-domain dialogues and community negotiations regarding data and information are developing. This report explains informatics and its place at the intersection of environmental, information, and social sciences as the participants in these fields collaborate in working with data that describe complex, interdependent environmental systems.

1.2 Working Definitions

Data management involves data handling (organizing, classifying, storing, labeling, and cross-indexing) and control (data definition, requirements, and quality) in order to make data available for use.

Information management involves the use of design, management, mediation, and communication principles for development of information systems as well as coordination of data flows that take into account distributed networks and long-term approaches to data access, preservation and exchange represented via data conventions, guidelines, protocols, standards, and best practices.

Informatics is the science of gathering, classifying, manipulating, modeling, preserving, and presenting, as well as designing for and teaching about data and information. It is the application of information science including information technologies and of social science including collaborative techniques to a field of study in ways that promote organization and flow of data, that highlight design of information systems and environments, and that draw upon concepts and methods to support data stewardship as well as the needs of data producers and users.

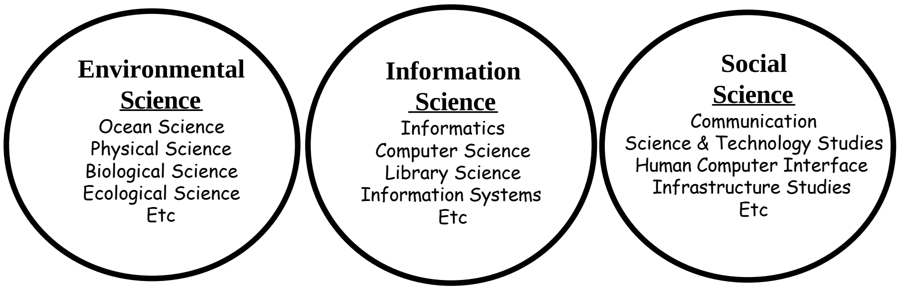

Figure 1.1. Related scientific fields of study grouped into environmental science, information science, and social science domains for the purpose of discussion. Each discipline or field of study within the domains handles issues of heterogeneity with data, data systems and data practices.

Consider three scientific research areas or domain sciences: environmental science, information science, and social science. Figure 1.1 shows these three sciences together with some of their constituent fields of study. Each of these scientific domains deals with patterns as well as with irregularities and heterogeneity. At any integrative scale (local, regional, global, temporal, organizational), the observational sciences deal with both homogeneous and heterogeneous data, standard and exceptional (heterogeneous) systems, as well as normal and atypical (heterogeneous) data practices. Critical to a full understanding of informatics is the recognition that heterogeneity and diversity can be valued as unique characteristics or problematized as irregular features.

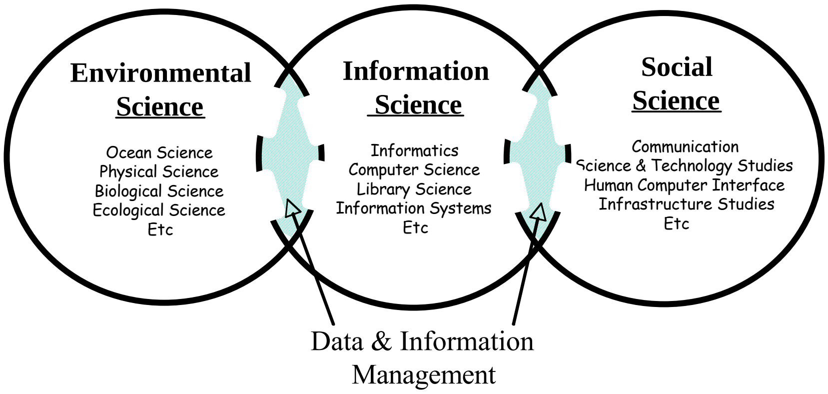

Figure 1.2. Data management and information management are illustrated occuring in overlapping domain areas (hatched). On the left an area of work is shown in the overlap between environmental science and information science while on the right a similar area of work is shared between information science and social science.

The work of data and information management occur partially at the overlap of information science and social science (Figure 1.2). Information science encompasses all the power of computer science and information technology while social science provides insight into human systems and interfaces, communication and design. In addition, data and information management occur partially at the interface between environmental science and information science (Figure 1.2b). The environmental sciences, ranging from atmospheric to ocean sciences and from geological to ecological sciences, are anchored by observational field and laboratory studies that produce data representing the earth and its processes. Within each of the sciences, including informatics, there are both field and theoretical components.

The intersection of environmental science, information science, and social science domains defines the arena of informatics (Figure 1.3). The work of informatics occurs at this Venn diagram intersection (hatched area), a center of interdisciplinary efforts. Informatics focuses on data and information processes that include collection, preservation, access, and use by local and distributed communities across spatial, temporal, and organizational scales. Informatics, like information systems in particular or science in general is a multi-faced endeavor integrating diverse techniques, protocols, and models.

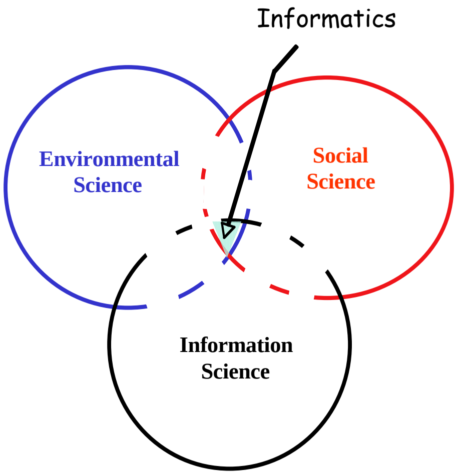

Figure 1.3. Environmental informatics framework: The work of informatics (hatched area) is shown at the intersection of environmental science, social science, and information science. The dashed areas suggest cross-domain communication.

1.3 Data Stewardship

Data stewardship is the management of data and its context, from planning to creation, organization, preservation, curation, and/or loss, involving data capture and describing, transformation and networking, use and reusing. Stewardship may center on work in the field, laboratory, project, program, center, or network but extends to concern with the dynamics and interplay of the whole system.

Through sampling, measuring, observing, analyzing, recording – direct observations, human operated instruments as well as automated platforms - scientists are now being inundated with data. Data, produced in the course of scientific investigations, are a resource to be managed. Management involves choices regarding availability, costs, and integrity with respect to data quality and formalizing accountability. When data interests mature to include an understanding of the long-term, data stewardship becomes key[4] .

Data and information management centers can be seen as carrying out a large set of activities as part of a data process sometimes described as the data lifecycle. It involves mediation work - that is, formal and informal knowledge elicitation, articulation, translation, and negotiation. Defined by its local or situated aspect in diverse organizations such as library, business, government, and science, information management is today, practiced by an emerging cadre of information professionals. Information management work extends to community level concerns with data accessibility and reuse as well as with user or stakeholder interests and engagement. From a data stewardship perspective, 'data work' is a suite of interwoven tasks and evolving processes sometimes called a lifecycle, rather than an existing single-package solution.

Informatics, information infrastructure, and design are all key elements for data stewardship that involve the development of human capabilities for mediation of data processes and for facilitation of appropriate dialogue about data processes amidst expanding data expectations, responsibilities, and costs.

1.4 Data Heterogeneity and Interoperability

Data Heterogeneity refers to multiple differences: among data collections and parameters within a collection, among data recorded according to varying criteria in terms of sampling, procedures and precisions; among data available resulting from a diversity of data records, descriptions, and analyses as well as from a diversity of models for storage, access and exchange.

Interoperability refers to a state or the goal of ensuring that the systems, procedures and culture of an organization are managed in such a way as to maximize opportunities for combination, exchange and reuse of data and information. (see UKOLN)

Data Interoperability refers to data defined and stored in such a way that like types can be identified for comparison and/or combination; it entails community conventions and standards, syntactic and semantic agreements.

New data and new data types as well as legacy data all are brought to bear in current scientific research inquiries. The challenge of heterogeneous data sets is distinct from that of the sheer abundance of data. The earth is a complex, heterogeneous environment of abiotic and biotic elements including the humans who inhabit it. Heterogeneity both constitutes the robustness of earth living systems and creates challenges for representation of these systems by data collectors and through data collections. Just as "the map is not the territory", so the data are not the object or the process they represent. There is an increase in heterogeneity that stems from the observation process itself, in terms of choice of sampling, instruments and methods that create not only a large number of parameters but also a number of ways of measuring the same parameter. Further, heterogeneity is manifest in digital records and data models for storage, access, description and exchange. The introduction of heterogeneity by human practices – often in spite of defined procedures and systems – is an important element that anchors informatics to everyday work. Networks of collections, systems and centers enable cross-comparison of data efforts and are beginning to reveal the extensiveness of heterogeneity.

Data sets that can't be combined, that "don't talk to each other", may not be interoperable because the constituent parameters or variables were collected for different purposes and in different ways. Interoperability is a state, something toward which one strives. The concept guides work with multiple, diverse assemblages of heterogeneous data. Data interoperability refers to collections of parameters or data collections being both accessible and working together. Working together necessitates explicit documentation of content and context commonality and/or relationships.

Interoperability has been described in terms of five categories: technical, semantic, political/human, inter-community and international (UKOLN). Work is ongoing in all these categories; signs of development take the form of guidelines, lessons learned, conventions, standard operating procedures, and best practices as well as standards and frameworks for creating standards.

Technical data interoperability depends upon the structure of the data and data types, relating for instance to standards in storage formats or in exchange formats for automated systems. There is also systems interoperability - the computer and network technology used for collection, storage, access, and exchange – that varies at any moment of time as ell as over time (Griffin, 1997).

Semantic interoperability is concerned with describing measurements and their context since data are collected in different locales with different instruments using different calibration techniques or recorded in different units. Semantic interoperability includes keeping track in machine retrievable form of well-known local details, such as whether a temperature measurement is in air or in water and whether a biomass measure is for water column phytoplankton or terrestrial plants. Metadata is one method for providing context of data through the use of tags that designate standardized categories for descriptive text about the data.

Political/ Human interoperability is involved in decisions about resources more widely available and also has implications for organizations, their staff and end users. Inter-community interoperability considers the increasing need to require access to information from a wide range of sources and communities. International Interoperability arises in working with other countries and the multitude of variations in standards, communications, and languages as well as communication styles and common grounds.

1.5 Information Infrastructure

Information infrastructure consists of intertwined technical and social components with organizational and conceptual aspects involving individuals and community participants carrying out work - designing, implementing, using, maintaining and redesigning - that frequently addresses the interface of human and information systems.

Design refers in data work to the ability to create data models, systems and infrastructures as well as to work with the multiple perspectives of participants, practices, and standards. Both for individual and collaborative data efforts, design involves the theory and practice of identifying a purpose, planning a strategy, and analyzing results while arranging parts and developing end products such as a database schema or a web interface.

An understanding of data and the work of informatics rests upon the concept of infrastructure. The term “infrastructure” in common parlance is used to refer to the pervasive physical support systems of pipes, wires, and pathways. As a broad functional category, though, it also includes an array of services and support (Star and Bowker, 2002). Examples of community infrastructure include schools and fire departments while in the digital realm there are computational services, help desks, and data resources. Stretching to the network realm, the term “cyberinfrastructure” alludes to alignments of interacting computational and technological resources. The Atkins Report (2003) summarizes: “If infrastructure is needed for an industrial economy, cyberinfrastructure is required for a knowledge economy”; between the base technology and its use are intertwined arrangements of “enabling hardware, algorithms, software, communications, institutions, and personnel”.

Infrastructure is required to facilitate data coordination (data capture, preservation, and use) and to support data interoperability. Information infrastructure must function in conjunction with physical infrastructure, such as digital connectivity, computational power, and storage capacity. The design of a functional information infrastructure relevant to multiple scales is a grand challenge across all sciences. For instance, consider how many information infrastructure elements are called upon in two very different situations: a) accessing satellite data from a national archive for download and use in a local visualization application for the conduct of a global research question or b) converting an online audio file from ogg to wav to ipod format for download to a pda in order to keep abreast of a colleague's work via an audio cast. Though representing different aspects of data work, the answer to the question of how many elements are involved may seem "a few" when all is familiar and functional but in practice the answer is "a host" – a host of standards, systems, and people who can mediate framing *questions as well as answers* given the multiple, interdependent elements that with some slight change may result in a misalignment(s) within the system.

To elaborate upon and effectively open up the work of informatics, our approach includes developing understandings of, sensitivities to, and vocabularies for work with data and its processes that are embedded within information infrastructures. Information infrastructure, a shifting blend of configurations and capacities of technology, organization, and community, is portrayed in Figure 1.4 (Baker et al, 2005; Ribes et al, 2005). The interdependent, persistent presence of all three of these mutually informing elements as part of recursive processes of constitution is inherent to work with data, work involving dialogue, design and development as well as deployment and enactment. These intertwined elements represent an ongoing effort; they represent sets of choices that change and are negotiated over time. A reminder of the need to be constantly aligning or (re)constructing is wrapped into the term "infrastructuring". This active form of infrastructure serves as a reminder that infrastructure is not just a thing but rather a set of arrangements, negotiations, and alignments that is a continuing state in terms of maintenance and update (Star and Bowker, 2002; Karasti and Baker, 2004).

Figure 1.4. Information infrastructure is viewed as an interdependent, multidimensional understanding of and sensitivity to configurations of technology, organization, and community.

A series of committees and groups have provided reports[3] on infrastructure and cyberinfrastructure, on long-term data collection and curation, on collaboratories and e-Science. These literatures, touching on the issues of interdisciplinarity and long-term sustainability, frame and prompt the work of informatics. Our understanding and construction of information infrastructures is developing alongside insights into cognition and communication dealing with media and messages that occur within disciplines and between disciplines.

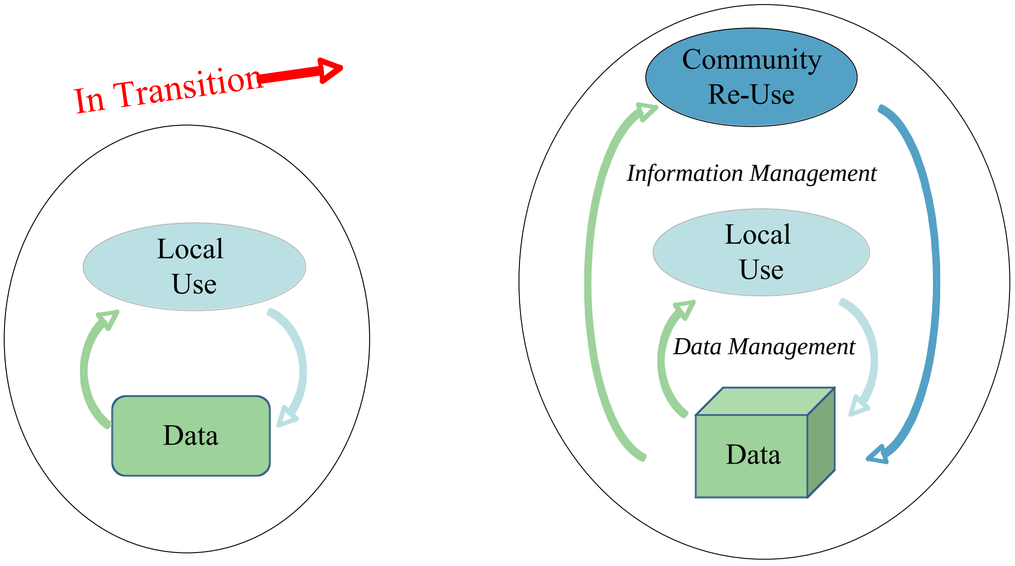

Figure 1.5. Information environments in transition: a) This model of local information infrastructure presents a traditional scenario involving data collection and the use of data for scientific research in a local arena; b) This model of community information infrastructure preserves the local information infrastructure involving data management efforts with collecting and local use of data but augments the model with information management efforts to establish community information infrastructure that supports data reuse over the long-term.

Our understanding of data, data collections, and data processes is also in transition. Community vocabularies and categories are only now beginning to emerge, often in response to changes in expectations with respect to data sharing and reuse. A traditional scenario for data collection has been to record data in the field either in hardcopy or digital format and to store this data in a central location: an individual's notebook, a disk holder on a bookshelf or on a group's laboratory computer. Such a local information infrastructure model is portrayed in Figure 1.5a.

In a scenario involving longer-term storage and wider-scale data use, Figure 1.5b shows capture and local use of data at the heart of a broader vision of infrastructure. Such an infrastructure model includes site-based data management and use in addition to larger-scale information management issues and longer-term use. Maintaining a system optimized for both local research and for distributed future use is an ongoing challenge.

Growth of effective information infrastructure helps to meet data challenges but often hinges upon design work. Design may refer to the practice of planning an element or a functional overview of elements (verb) as well as to an arrangement of parts or an end product such as a symmetric design or a design sketch (noun, adjective). Design work takes into account the interface of human and technical systems. Design approach and capacity are key to initiating informatics, infrastructuring, and information environments.

The literature provides growing evidence of the importance of design as both an approach and a tool. Design work is under discussion as engineering and user-centered design, participatory and ecological design. Emphasis on the design and standard-making processes occurs in the fields of information systems and action research and analytic ethnography[5]. Participatory Design, Computer Supported Cooperative Work, and User-Centered or Meta-Design[6] are contemporary approaches that allow us to consider how we do our work as we work. In informatics, the design process is one element of "infrastructuring".

2. New Challenges with Data Processes

2.1 Expanding Data Requirements

There are new requirements along with new data types, data functions, and participant roles associated with the data that are central to observational environmental science. Data requirements change depending upon needs of the scientific researchers, expectations of support agencies, and plans of programs or communities. At the individual or well-integrated group level, the collecting of field data has frequently been planned for immediate, specific scientific purpose. In such a case, the field team researcher or support staff is frequently the manager of the data. There is a tight link between measuring, recording, and using (Figure 2.1a). Today, with a transition underway in the work of data collecting, the notion of a single entity called 'data' has opened up into a view of data involving collecting and managing where data management is a distinct element in the data collection process (Figure 2.1b).

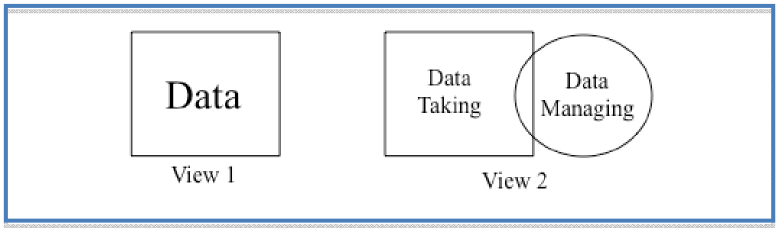

Figure 2.1. Two views of data collecting: a) As a single entity, an independent data effort and b) as an expanded view distinguishing data taking and data managing.

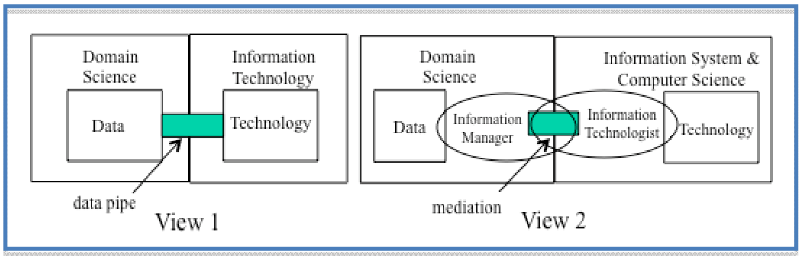

Figure 2.2. Two views of preserving and/or exchanging data showing the development of information management and information technologist roles: a) View 1 illustrates the concept of an automated data pipe and b) View 2 presents a bidirectional mediated interface with bidirectional communication between data source and data archive.

The data process may be discussed in greater detail using a simplistic multi-function model for illustration as summarized in Table 2.1. Five overlapping data functions with differences in focus and activity are identified: data taking, managing, information managing, data preserving, and data and information exchanging. The first function, data taking involves data collectors who may be a variety of participant types from researchers to technicians and technologists. Next, data managing focuses traditionally on local data use involving data managers but may grow to include an online data access and a site-based repository through the efforts of information managers and perhaps an informatics team. Data preservation represents a separate function, focusing on archiving supported by data curators and librarians. For this function, there may be some informal or project-guided activity aimed at data sharing along with attention to migration of artifacts to accommodate changing technologies and media. Finally, explicit data reuse is a fifth element of the data process that is operative in networked arenas, designed and supported by information professionals and in the best of circumstances guided by user needs from those of site scientists to those of community scientists, policy makers and the public. Data preservation for reuse moves beyond local data use so requires contextualization of the data through construction of metadata (data about the data) and for preparation of data in accord with standards for data interoperability and exchange.

Table 2.1 Elements in a Data Process Model

Element	Function	Focus	Activity	Associated Roles
1	Data Taking	Field	Sampling	Data Collectors & Data Technicians
2	Data Managing	Site-Based	Use	Data Managers
3	Information Managing	Site-Based	Repository	Information Managers & Informatics Team
4	Data Preserving	Long-Term	Archive	Data Curators & Librarians
5	Data Exchanging	Network	Reuse	Information Analysts & Technologists

Roles associated with expanded requirements in the data flow process are emergent and necessarily overlapping; the role of information management was shown earlier at the interface of domains that generate and handle data. Information management includes facilitating the data flow among multiple functions (Table 2.1). Adding preservation and exchange of data produces an under-appreciated nonlinear scaling of the work associated with the data process. Though included in larger-scale organizational information architecture plans, individuals, groups, and institutions are only beginning to recognize and take into account these functions and their activities when planning data efforts.

While views of data collecting are in transition, techniques for data preservation and exchange are in formation, being worked on by a diverse spectrum of participants such as environmental researchers and technicians as well as managers, technologists, library scientists and computer scientists. When technology is brought to bear on data issues, perceptions of the data as well as the data process itself change. Frequently, a data pipe is imagined connecting data to a black-box technology (Figure 2.2a). For an observational scientist intent on maintaining a field research program, there is a hope or expectation that this data pipe exists as an automated, low maintenance direct connect to a well-defined and bounded data management solution.

Data exchange and preservation are relatively new activities. To design and carry out this work, new roles are emerging such as the data or information manager (IM) mediating between the environmental scientist and their data as well as the information technologist (IT) mediating between data and information technology (Figure 2.2b). A single individual sometimes performs both roles. Involving more people and maintaining communication among them is central to contemporary data care. Informatics brings an opening-up or improved transparency that expands upon or replaces notions of data pipes and technological black boxes.

2.2 Expanding Data Responsibilities

In addition to the roles emerging in the data handling process, there are expanding responsibilities for those who participate in observational scientific research and informatics. Zooming in on any part of the data life cycle reveals an array of tasks and relations with distinct data responsibilities.

Though an environmental scientist must get on with the work of environmental science, information management and information infrastructure are a growing part of their science. With the development of more complex scientific questions, tools, and information sources, project management necessarily involves planning and management of both field data and information resources. A range of nested efforts from local disk storage requirements to community data federation strategies and organizational infrastructure building projects to participation in national cyberinfrastructure and standard-making endeavors are all part of planning efforts. Indeed, scientists and information managers are being tasked with handling vast amounts of new data as well as with creating new policies, interfaces, associations and relationships.

The work of information managers involves managing data and designing information systems as well as the growth of information infrastructure and the professionalization of their role (Baker and Karasti, 2003). Three distinct areas of interface have been identified as part of long-term information management: science, data, and technology (Figure 2.3a; from Karasti and Baker, 2004). Information managers resolve tensions and create a balance among these interdependent elements. Their work draws on diverse fields, local knowledge and working experience. On a day-to-day basis work includes mediation within multiple timeframes: the short-term (interrupt driven tasks), the mid-term (negotiated products) and the long-term (maintenance and (re)design of scientific and infrastructure processes).

Global questions and networked science bring expanded roles with new responsibilities to scientists as well (Figure 2.3b). A traditional model for an observational earth scientist involves taking data and then using data and presenting findings. As scientific endeavors have grown to become multi-institutional and longer term, establishing and maintaining effective scientific programs and communities has become a major component of the work of the scientist. In many cases, the single technical field assistant and shelf of notebooks is being replaced or augmented by a community team using a plethora of physical and digital infrastructure elements as work with data has burgeoned both for immediate personal use as well as for future community uses.

2.2.1 Mediation

Modern data responsibilities of domain scientists and information managers involve a strong component of mediation work. Mediation work differs across the disciplines but is required to interface and to create new combinations of datasets and organizational systems with differing foci and activities. Mediation involves articulation and translation work in order to negotiate, interface, and align requirements in the design of information flows among data taking, management, archive, and exchange activities, among data types and structures, as well as among data handlers of all types.

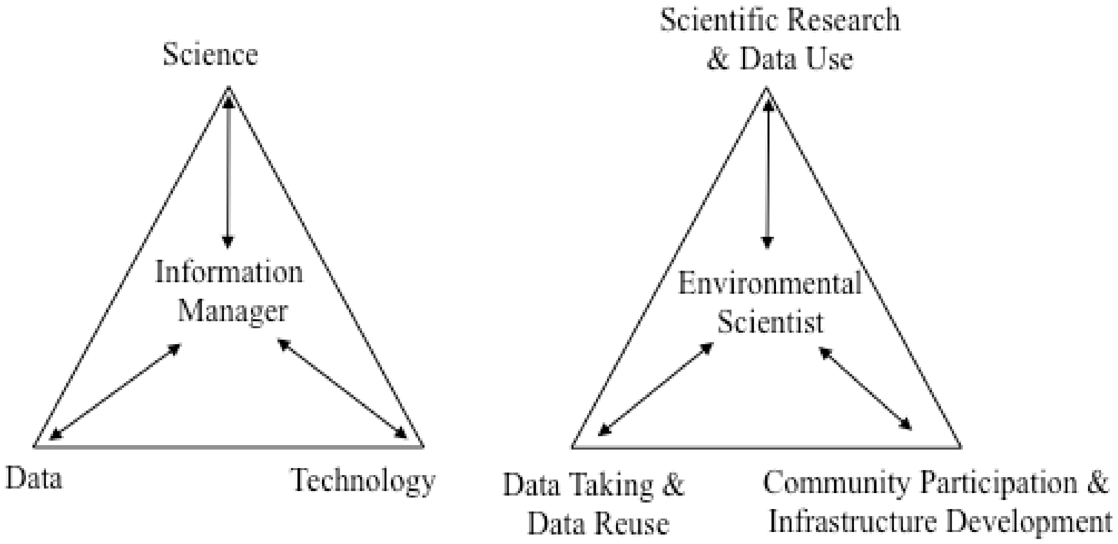

Figure 2.3. The roles of information manager and of environmental scientist have expanding responsibilities in terms of balancing multiple elements: a) Relationships between the role of information management and responsibilities with science, data, and technology (from Karasti and Baker, 2004) and b) Relationships between the role of environmental scientist and responsibilities with scientific research and data use, data taking and data reuse as well as community participation and infrastructure growth.

Frequently roles may be misaligned or may be construed differently. Figure 2.4 provides one example of two different perspectives on the work of mediators. From a domain science perspective that includes the notion of data pipes, there is a view that data tasks are problems that can be solved technologically where technology is the province of information managers and technologists, closer to the realm of information system and computer scientists (Figure 2.4a). On the other hand, from a computer science perspective, the work of the information technologist represents a field endeavor closer to the fieldwork of domain scientists (Figure 2.4b). These two views come together in Figure 2.2b to make evident that the practices of one influences the practices of the others. This is interdisciplinarity with all its attendant mediation.

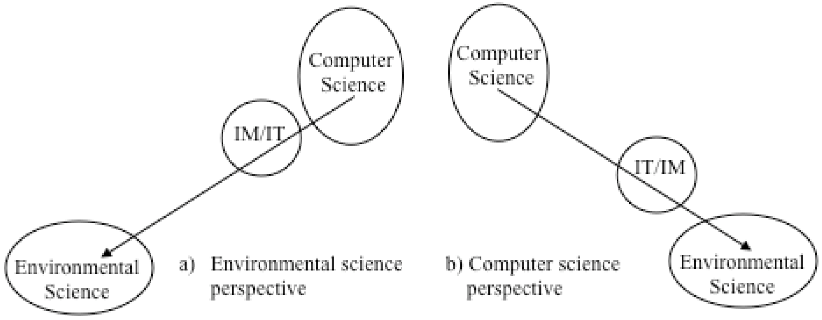

Figure 2.4. Two perspective on mediation roles: a) an environmental science point of view where information management and information technology are considered close to information system development and computer science and b) a computer science point of view where information technology and information management are perceived as closer to environmental science.

2.2.2 Metadata

Describing the data and the context of data involves selecting, within a sometimes large, divergent and even discordant scientific community, metadata categories. Such categories have ramifications in terms of data findability, access, and interoperability. For example, controlled vocabularies, dictionaries, taxonomies, and thesauri create categories for data management so that data entries such as ‘oxygen’ and ‘O2’ can be related

Another example involves the existence of distinct data types within different research groups and projects such as ‘streamed data’ and ‘manually-handled data’. Distinct data types involve differing digital foci, in this case an emphasis on real-time automated homogeneous data acquisition versus hypothesis-driven heterogeneous hands-on data collection. Larger or broader scale data endeavors incorporating information systems have technology techniques and data organization formalized in data schema that document particular relations and enable data access and data query.

2.2.3 Interoperability

Creating and maintaining data interoperability involves diverse arrangements and actions, including coordinating the data process and the people who handle data from the moment of capture to that of use. Standards are a non-trivial element of interoperability and for the most part do not exist, have not been enacted, or have changed at the local level. At the site or local level, a research group may not yet have identified which of the data characteristics are important to document in order that the data can be of use by others working in widely differing contexts. Standards that have been developed, agreed upon, and enacted are examples of collaborative work enabling interoperability. Scientific communities are currently in the process of learning how to go about developing and enacting agreed upon standards that capture the characteristics of data at hand. Rather than establishing rules to enforce, standard-making is a new form of work that requires new types of design, learning, and training for research community participants as well as user populations. Though our understanding of informatics and information roles is undergoing development and our discursive capacity expanding, it remains difficult to acknowledge and fund such work. Dialogue and forums are needed to bring together those in diverse expert communities for identifying existing practices, building consensus and prototyping standards.

2.3 Expanding Data Costs

The most grievous effect of data becoming inaccessible or unusable for research is, of course, the loss of knowledge that the data could have engendered. But there is also a very considerable associated financial loss incurred when such data are “lost.” Figure 2.5. The digital costs for collecting, storing and accessing data over time. The far right column represents a cumulative total.

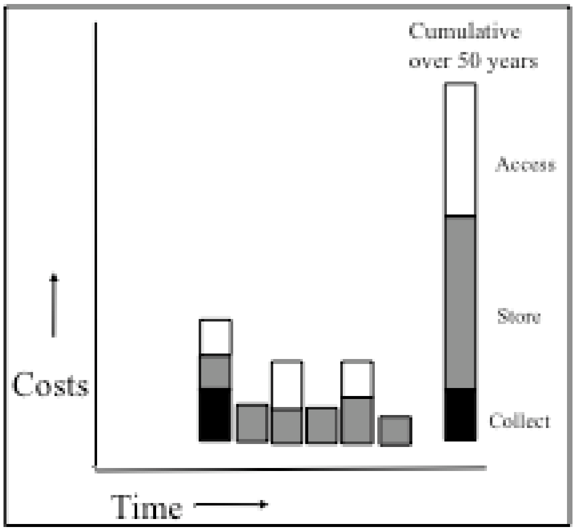

Figure 2.5. The digital costs for collecting, storing, and accessing data over time. The far right column represents a cumulative total.

A typical data scenario involves a field research scientist awarded support to take data and then analyze and publish research results. Traditionally, the cost of collection has dominated data budgets. When data preservation, publication, and exchange are added, however, new costs are introduced including annual storage, continuing curation, and periodic updates to hardware and software (Figure 2.5; from Baker and Bowker, in press). If access is lost, the data and many years of collection and analysis investment are lost as well.

New categories of data costs must be considered carefully both in terms of local knowledge work and site-based curation of data repositories as well as in terms of community coordination and national archives. We have yet to experience or to take into account fully the ramifications of many of our strategic choices. Scientists have long faced the issue of selecting data to be collected but now face questions within the informatics arena regarding what data to select for preservation over long time periods. Such considerations are best initiated with data collection discussions during the design phase of fieldwork. Field experiments have worked within the bounds of limited resources in terms of equipment and field time as well as analytic and human resources. To this must be added problem formulation and data scoping for the handling and analysis in terms of time for information systems support, data documentation, and user interfaces.

3. Ocean Informatics

Ocean Informatics is the application of informatics to the domain of Ocean Science. It is the work that occurs at the intersection of oceanography, science studies, and information management.

With a focus on marine science, Figure 3.1 transforms a general use of the term informatics (Figure 1.3) by specifying the environmental science as Oceanography, the social science as the field of Science Studies, and the information science as Information Management. The intersection represents our local situation and is designated Ocean Informatics (Baker et al, 2005).

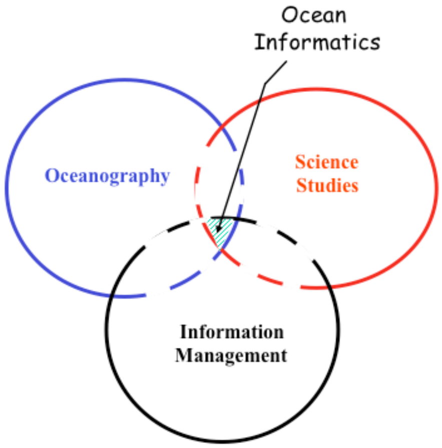

Figure 3.1. Ocean Informatics: Our local configuration for ocean informatics is shown (hatched area) at the union of Oceanography, Science Studies, and Information Management.

Ocean Informatics is a conceptual framework bringing together theory and practice for those working with oceanographic data, and promotes an ongoing effort drawing on social theory and the principles of information science. Informed by a broader context, local design is grounded by data taking and use data. At Scripps Institution of Oceanography, our ongoing interdisciplinary data effort (http://oceaninformatics.ucsd.edu) represents just one of the many site-based efforts required to create an institutional information management strategy.

3.1 Heterogeneous Ocean Data, Systems, and Networks

Many of the challenges for oceanographic information management and marine science infrastructure design arise from the very nature of oceans. When the field of study is oceanography, the object of study is the global ocean, in and of itself, a vast medium; its surface covers more than 70% of Earth's surface while its greatest depths exceed the dimensions of Mt. Everest.

But the ocean is never "in and of itself" because the waters of the global ocean interact with the planetary atmosphere, land forms and deep-earth tectonic activity. To study the ocean is to study the abundant plant and animal life found there, the physical, chemical, geological, and atmospheric interfaces of the water and the myriad interactions and mutual influences of all these aspects of the global ocean system. To study the ocean is to work on vast spatial and temporal scales that reach back into the Earth's earliest stages of evolution and forward into the future we are creating today.

Sampling, measuring, observing, recording - with instruments borne on ships, satellites, moored buoys, drifting buoys, and remotely operated platforms - oceanographers are inundated with data. The following scenario provides a classic illustration of a data discussion in the present digital, scientific age:

> *"Mia, over here." Don rearranged the laptop and papers on the table so the approaching woman could put down her coffee cup.*
>
> *"I sure hope you had better luck than I did, Don." was her greeting. Mia, an oceanographer, had felt the task to be seemingly straightforward: assemble characteristics for a comparative study using biomass data as an indicator of water column production. She and Don, the research team data manager, had worked over the last decades with four easily identified datasets: two Antarctic circumpolar current (ACC) datasets consisting of the historic British Antarctic Survey Discovery data and the Palmer LTER time series annual cruise data; two eastern boundary current (EBC) datasets consisting of the CalCOFI annual time series cruise data and the California Coastal Studies Data Zoo collection. They had divided up the task of obtaining data: Don would get the ACC data and Mia, the EBC data.*
>
> *"Well," Don replied, "I did find the Discovery data, but it was used a decade ago for another project and stored on a floppy from an old HP9845 that has long since gone to cyber heaven, making the disk useless. Oh, I eventually found the data; guess where, in carbon copies of the summary sheets made earlier from reports stashed away on library shelves. It only took me two days to enter and check the data." Mia grimaced in sympathy, "What about the Palmer data; that's online already, right?" "Yes, Don acknowledged, "but the site information manager says her group is in the midst of migrating the data from individual online files into time-series datasets complete with ecological metadata. Brace yourself – many more months before that's finished and accessible. Could you track down the CCS web site?"*
>
> *"That was the easy part. Some of the data is in NetCDF format so we have to gather the applications online for parsing the data. Fortunately, the CalCOFI web page was easy to google, but unfortunately, to group the bottle and continuous profile data, I will have to download and install a beta version of the CIFT Windows 9x/NT program."*
>
> *"Mia, did you want to work with all four datasets using our Matlab visualization scripts? I just talked with Jerry, the department systems manager; he mentioned that the PC with the Matlab software crashed yesterday and they will install Matlab on their secondary workstation 'as soon*

> *as they have time.' And…remember, I mentioned before, that Jerry is in the midst of renegotiating department's site license for the updated version of Matlab"*
>
> *"And we thought we were doing data aggregation," Mia sighed, "This feels more like data aggravation."*

Here discontinuities and misalignments are evident that have immediate ramifications for scientific work. Oceanographic data management work is indeed situated within a highly interdisciplinary arena of heterogeneous data, formats, systems, and networks.

3.2 Informatics Good Practices

Information is plentiful and pervasive especially when augmented by technological capabilities. There is simultaneously work to be done and frameworks to be developed in order to manage data and their supporting information infrastructures. Much exciting work is opening up in informatics, in partnership with other domains. Yet there are fundamental unresolved issues of scale and relations even as computational capability expands from petaflops to exaflops, as network rates increase from megabits per second to gigabits per second, and storage grows from gigabytes to petabytes. Informatics issues and their ramifications range across a wide set of goals (resource use, preservation, quality, learning) and models (data types, access, exchange, responsibilities) to the realms of human values (information power, control, risk, policy) in general and ethics (data sharing, technological ramifications, sustainability, and well-being) in particular. The interdependence of human and technological systems carries with it the need for a keen awareness of potential long-term ramifications - and unintended consequences.

To address the multitude of data issues, new interdisciplinary approaches must be invoked to work with current and legacy technologies, to develop standards and distributed networks as well as to identify and articulate data processes that enable design, development, deployment, and enactment of information systems. This entails re-examining the nature of informatics work, its practices, its practitioners, and its interface with technology. Such work is framed by our choices with respect to representation, participation, and science but inevitably these choices depend on the choices we are able to *see* both short-term and long-term. Informatics work is made more challenging at present as there are few local support mechanisms in place either for doing broader-scale and/or large-scale programmatic science in new ways or for crossing the boundaries of existing structures, cultures, and traditions. Ultimately, because scaling is possible, creating mechanisms and processes so that participants can weigh in on the balance of leveraging existing configurations and establishing alternative arrangements is an important aspect of design.

Table 3.1 Informatics Good Practices

1. Incorporate data problem formulation and data scoping early in the scientific planning process.
2. Recognize articulation, translation, negotiation and mediation as central to work with data.
3. Partner with appropriate information professionals for data work.
4. Create collaboration opportunities as well as coordination mechanisms for community work.
5. Recognize informatics as conducting research while carrying out information management.

Given our emergent understanding of data and informatics, a series of summative overviews - by no means exhaustive - are brought together for joint consideration (see Appendix 7.2). Selected summaries serve as an aid for those interested in a broader exposure to principles, gestures, elements, findings, and strategies. Some of our experiences with Ocean Informatics data stewardship, a kind of preliminary lessons learned, are summarized below (Table 3.1).

These five points have emerged in informatics work carried out in our information management-environmental science partnerships; they highlight the need for long-term planning and community-scale collaborations. To the extent that these suggestions are understood and incorporated into everyday work, they may contribute to an escape from today's quasi-permanent state of over-commitment to status quo of short-term scientific pursuits and to traditional data practices.

3.2.1 Incorporate data problem formulation and data scoping early in scientific planning process.

Data scoping as practiced today for individual, project, and organizational efforts is inadequate for longer term, larger scale endeavors. Long-term ramifications are under-appreciated at a time there is developing a new understanding and articulation of the work involved not by information managers, technologists, or scientists alone but as a co-informed, mutually coordinated effort. Data scoping involves the ability to recognize and arrange time to consider data work in full, taking into account the data life cycle and multiple time scales, technical environments as well as organization and community factors.

3.2.2 Recognize articulation, translation, negotiation and mediation as central to work with data.

Mediation work focuses on data and technology and their relations with a particular science task. Mediation facilitates formulation of informatics issues as having social, technical, and organizational components to be addressed as design challenges rather than designating or delegating data management as a problem that can be solved by technology alone. Delegating - similar to outsourcing - creates a series of often-unrecognized consequences for local data efforts that cost more time, upset, and risk when they reappear subsequently as legacy issues. Many of these issues are negotiated and addressed when the design process is both recognized and supported.

3.2.3 Partner with appropriate information professionals for data work.

Problem formulation, data scoping and informatics work in general require both involvement and engagement of site data and information managers early in the planning. Representation or planning of longer term, broader scale data work by non-specialists is problematic, often resulting in subsequent data recovery efforts that are distinctly reactive rather than proactive. Environmental scientists are absorbed by changing inquiry scenarios and growing expectations within their own fields while the approaches of information technologists tend to emphasize technological solutions germane to technology and computer science rather than to local science needs. Engaging the informatics community up front in defining the work collectively means mediation and local learning are not delayed or forgotten. Information professionals are able to negotiate within collaborative arenas and to face a host of inherent dichotomies - local and global requirements, short-term and long-term plans, technical and social approaches, practical and theoretical design, and so forth – by recognizing that reframing these potential tensions is key to enabling reciprocal relations and learning opportunities.

3.2.4 Create collaboration opportunities as well as coordination mechanisms for community work.

Collaborative techniques are required to support joint work from proposal writing to systems design, particularly as a strategy to foster communication, integration, and innovation. Take the example of joint work simplified into steps: a) discuss priorities, contexts, and their interrelations; b) assemble and synthesize elements; c) identify a storyline; d) express ideas in writing. Traditionally a unified research view is achieved by conducting a group discussion (step a) prior to an individual taking the oral to print by performing steps b-d. In contrast, group cooperation during these latter steps is what transforms the work of polyphonic blending into an integrative step where potential for discovery of an emergent view is high. This takes more planning and time than is traditional yet in this manner a proposal writing session becomes a learning activity for multiple participants in contrast to a single individual's heroic effort at smoothing and integrating a collection of exchanged lists. Moving beyond coordinated work to collaborative work as well as collaborative design means seeing beyond the short-term goal of getting the work done to fostering participatory learning opportunities and considering how the work is done. Such a concern has been called out as 'collaborative care' (Jackson and Baker, 2004). One ramification of this approach is that it effectively disperses authority and responsibility while engaging interest of participants.

3.2.5 Recognize informatics as conducting research while carrying out information management.

Informatics is emerging organizationally as a field of scholarship; its interdisciplinary nature spans existing academic categories and configurations. Ongoing changes in data requirements and expectations (such as data sharing, data reuse, and network participation) for environmental scientists give impetus to move from framing information management work as solely technical support to recognition of design as representing a blend of both application and research endeavors. Informatics draws on the concepts of collaborative design and participatory learning as transformative and self-renewing approaches providing an alternative and an impetus for new types of data use and of scientific collaboration.

4. Informatics is Happening

Our notion of informatics is indeed "in formation". Our understanding of data, knowledge, and information as part of the work of informatics in particular and information science in general is also under development. Today, informatics may be viewed as a field of study or as a scientific discipline. Within the academic arena, information work appears under many names and is emerging in many organizational forms, as departments, centers and schools (see Appendix 7.3).

The work of creating an information infrastructure across multiple dimensions (Figure 1.4) amidst transitioning models and timeframes (Figures 1.5) is a complex endeavor and necessarily draws upon insights from a number of sciences (Figure 1.1). Informatics integrates across scientific domains (Figure 1.3) and helps achieve new perspectives that are required to create and maintain contemporary information infrastructures.

There are significant issues in the field of informatics that require further research. These are not esoteric matters but have a multitude of ramifications in everyday work; they are a product of the many types of representation involved in the scientific work of organizing, describing, classifying, preserving, and making available data. In this work tactical and strategic choices introduce limitations in addition to benefits in a manner similar to the situation faced by scientists in planning field sampling that portrays a selected portion of a natural system,

Today, data interoperability and information infrastructure are integral to environmental science research. This report has presented how informatics involves technology, organization, and people in communities with data and knowledge practices. People organize in study groups and university departments, in local institutions and national centers, in trans-national organizations and networks of networks. People are part of communities of interest and communities of practice, physical communities and virtual communities. Their work involves interfaces between individuals and groups, activities and disciplines, communities and organizations. It is people who handle data and align infrastructure elements, choose categories, make relations between entities and then know the appropriate moment to change arrangements and redesign. The work of informatics professionals involves balancing and juggling between spatial and temporal scales, between mediation and implementation work.

Amidst such open-system complexities, informatics brings with it a concern and conceptual framework for data stewardship inclusive of an awareness of the multiplicity of data processes and a full data lifecycle. Informatics – when conducted with an awareness of its multiple dimensions - together with data stewardship hold the potential to enhance our human capacity to envision and understand environmental, human, and information systems and their relations as a whole earth ecosystem.

5. Acknowledgement

Recognition goes to the Ocean Informatics team members and their contributions to collaborative design of a new type information environment. Our organizational setting is the Integrative Oceanography Division, Scripps Institution of Oceanography (IOD/SIO) at the University of California, San Diego and the community umbrella provided by the Long-Term Ecological Research (LTER) Program as well as the California Cooperative Oceanic Fisheries Investigation (CalCOFI) program. Thanks and acknowledgement to Dawn Rawls for scientific editing. This work was initiated with early support from National Science Foundation Grants NSF/EIA-01-31958 and NSF/DBI-01-11544 and developed under support from across NSF directorates: NSF/OPP-02-17282, SBE/SES 04-33369, and NSF/OCE NSF 04-17616.

6. References and Reference Notes

References

Atkins, D. et al., 2003. NSF-AP Report: Revolutionizing Science and Engineering Through Cyberinfrastructure: report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. http://www.communitytechnology.org/nsf_ci_report

Baker, K.S. and G.C. Bowker, Information Ecology: Open System Environment for Data, Memories and Knowing. Journal of Intelligent Information Systems. BDEI Special Series, (forthcoming)

Baker, K.S., S.J. Jackson, and J.R. Wanetick, 2005. Strategies Supporting Heterogeneous Data and Interdisciplinary Collaboration: Towards an Ocean Informatics Environment. Proceedings of the Hawaii International Conference for System Science (HICSS38).

Baker, K.S., D. Ribes, F. Millerand, G.C. Bowker, 2005. Interoperability Strategies for Scientific Cyberinfrastructure: Research Practice. Proceedings American Society for Information Systems and Technology

Berman, F. and H. Brady, 2005. NSF SBE-CISE Workshop on Cyberinfrastructure and the Social Sciences, Final Report. URL http://director.sdsc.edu/pubs/SBE/reports/SBE-CISE-FINAL.pdf

Birnholtz, J. and M. Bietz, 2003. Data at Work: Supporting Sharing in Science and Engineering.

Black, A. and R. Brunt, 1999. Information Management in Business, Libraries and British Military Intelligence: Towards a History of Information Management. Journal of Documentation. 55(4): 361-374.

BMIS British Medical Informatics Society (http://www.bmis.org)

Bowker, G.C. and S.L.Star, 2000. Sorting Things Out: Classification and Its Consequences. The MIT Press.

Bowker, G.C, 2001. Biodiversity Datadiversity, Social Studies of Science, 30(5), 643-684.

Brunsson, N. and B. Jacobsson and Associates. 2000. A world of Standards. Oxford, Oxford University Press.

Brunt, J.W. P. McCartney, K.S. Baker and S. Stafford, 2002. The Future of Ecoinformatics in Long Term Ecological Research, in 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.

Chae, B. and M.S.Poole, 2005. The surface of emergence in systems development: agency, institutions, and large-scale information systems. European Journal of Information Systems 14:19-36.

Chen, W. and Hirschheim, R., 2004. A paradigmatic and methodological examination of information systems research from 1991 to 2001. Information Systems Journal 14(3): 197-234.

Cherns, A., 1976. Principles of sociotechnical design. Human Relations 29(8), 783-792.

Cherns, A., 1987. Principles of sociotechnical design revisited. Human Relations 40: 153-162.

Choo, C.W., 2002. Information Management for the Intelligent Organisation: the art of scanning the environment, Medford, NJ, Information Today.

Clegg, C.W., 2000. Sociotechnical Principles for System Design. Applied Ergonomics 31:463-477.

David, P., 2005. Towards a Cyberinfrastructure for Enhanced Scientific Collaboration: Providing its ‘Soft’ Foundations May be the Hardest Part. Oxford Internet Institute, Research Report No. 4.

Dittrich, Y., S.Eriksen and C.Hansson, 2002. PD in the Wild: Evolving Practices of Design in Use. In Binder, T., J.Gregory and I.Wagner (Eds) Proceedings of the Participatory Design Conference Malmo, Sweden, 23-25 June 2002. CPSR, Palo Alto, CA, 2002.

Department of the Navy, 1999. http://www.chips.navy.mil/archives/00_jul/data_interoperability.html

Edwards, P., 2003. Infrastructure and Modernity: Force, Time, and Social Organization. In T.J. Misa, P. Brey and A. Feenberg. The History of Sociotechnical Systems. Modernity and Technology. Cambridge, MA, The MIT Press.

Finholt, T., 2003. Collaboratories as a New Form of Scientific Organization. Econ. Innov. New Techn 12(1): 5-25.

Finholt, T., 2004. Collaboratories. In Annual Review of Information Science and Technology. E.B. Cronin (ed) 36: 73-107. Medford NJ

Fischer, G., 2003. Meta-Design: Beyond User-Centered and Participatory Design. HCI Proceedings URL: http://www.cs.colorado.edu/~gerhard/papers/hci2003-meta-design.pdf

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

Fujimura, J.H. 1992. Crafting Science: Standardized Packages, Boundary Objects, and Translation. In Science as Practice and Cutlrue, edited by A.Picering 168-214. Chicago, University of Chicago Press.

Futrell, J. and AC-ERE, 2003. Environmental CyberInfrastructure: Tools for the Study of Complex Environmental Systems. AC-ERE. http://www.nsf.gov/ere

Granger, S., 2002.Digital Preservation and Deep Infrastructure. D-Lib Magazine 8(2). URL: http://www.dlib.org/dlib/february02/granger/02granger.html

Graves, S., C.A.Knoblock, and L.Lannom, 2002. Technology Requirements for Information Management, RIACS Technical Report 2.07, November 2002, http://www.riacs.edu/navroot/Research/TR_pdf/TR_02.07.pdf

Greenbaum, J. and M. Kyng (eds), 1991. Design at Work. Lawrence Erlbaum, Hillsdale.

Griffin, S. M. Digital libraries and the NSF/DARPA/NASA Digital Libraries Initiative. In: David Raitt (ed.) Libraries for the new millennium. London: Library Association Publishing, 1997, pp 114-147.

Hedstrom, M. et al (eds), 2002. NSF-DGO It's about Time: Research Challenges in Digital Archiving and Long-Term Preservation. Report on NSF Workshop on research challenges in digital archiving

Hey, T. and A. Trefethen, 2003. The Data Deluge: An e-Science Perspective. in Berman, F., Fox, G.C. and Hey, A.J.G., eds. Grid Computing - Making the Global Infrastructure a Reality, Wiley and Sons, pp. 809-824.

Iivari, J., 1991. Paradigmatic Analysis of Contemporary Schools of IS Development, European Journal of Information Systems, Vol. 1-4, pp. 49-272.

Jackson, S.J. and K.S. Baker, 2004. Ecological Design, Collaborative Care, and Ocean Informatics in Proceedings of the Participatory Design Conference, 27-31 July, Toronto, Canada, CPSR, Palo Alto, CA.

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

Karasti, H., 1994. Increasing Sensitivity Towards Everyday Work Practice in System Design. PhD Thesis. Department of Information Processing Science, University of Oulu, 2001.

Karasti, H. and K.S. Baker, 2004. Infrastructuring for the Long-Term: Ecological Information Management. In Hawaii International Conference for System Science Proceedings, Big Island, Hawaii 5-8 Jan 2004.

Karasti, H., K.S. Baker, and E. Halkola, forthcoming. Enriching the Notion of Data Curation in e-Science: Data Managing and Information Infrastructuring in the Long Term Ecological Research (LTER) Network. In M. Jirotka, R. Procter, T. Rodden, and G. Bowker (eds). Computer Supported Cooperative Work: An International Journal. Special Issue: Collaboration in e-Research

Khazanchi, D. and B.E. Munkvold, 2002. On the rhetoric and relevance of Information Systems Research Paradigms: A conceptual framework and some propositions. Proceedings of the 36th Hawaii International Conference on System Sciences, IEEE Computer Society Press.

Kling, R., 2000. Learning About Information Technologies and Social Change: The Contribtuion of Social Informatics. The Information Society 16: 217-232.

Lord, P. and A. Macdonald, 2003. e-Science Curation Report.; JISC Committee for the Support of Research. www.jisc.ac.uk/e-sciencecurationreport.pdf

Lyytinen, K., 1987. Different perspectives on information systems: problems and solutions. ACM Computing Surveys 19(1): 5-46.

Macevičiūtė, E. and T.D. Wilson, 2005.□The Development of the Information Management Research Area. in: Elena Macevičiūtė and T.D. Wilson, eds., Introducing Information Management: an Information Research Reader, pp. 18-30. London: Facet Publishing.

MacLean, M. and B.H. Davis (eds), 1998. Time & Bits: Managing Digital Continuity. J.Paul Getty Trust, Los Angeles.

Michener, W.K., J.W.Brunt, and S.G. Stafford, 1994. Environmental Information Management and Analysis: Ecosystem to Global Scales. Taylor and Francis, London.

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

Morton, M.S., 1996. How Information Technologies Can Transform Organizations. In Computerization and Controversy, Rob Kling (ed), Academic Press, New York, p 148-160.

NRC, 1995. Finding the Forest in the Trees: The Challenge of Combining Diverse Environmental Data: Selected Case Studies.

NSB National Science Board, 2005. Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century http://www.nsf.gov/pubs/2005/nsb0540/start.htm

NSF Report, 2003. Knowledge Lost in Information. Report of the NSF Workshop on Research Directions for Digital Libraries. http://www.sis.pitt.edu/~dlwkshop/report.pdf

OED, Oxford English Dictionary 2nd Ed., 1981. Oxford University Press, Oxford.

Parsons, M. and R.Duerr, 2005. Designating User Communities for Scientific Data: Challenges and Solutions. Data Science Journal 4:31-38.

Penniman, D., 2002. What Exactly is Informatics?. University at Buffalo Reporter 33(23). URL http://www.buffalo.edu/reporter/vol33/vol33n23/pa.html

Pfirman, S., 2003. Complex Environmental Systems; Synthesis for Earth, Life and Society in the 21st Century. A report summarizing a 10-year Outlook for the National Science Foundation.

Raymond, E.S., 1999. The Cathedral and the Bazaar. Cambridge, O'Reilly.

Ribes, D.. K.S. Baker, F. Millerand, G.C. Bowker, 2005. Comparative Interoperability Project: Configurations of Community, Technology, and Organization. Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries. New York, ACM Press.

Sawyer, S., 2005. Social Informatics: Overview, Principles and Opportunities. Bulletin of the American Society for Information Science and Technology 31(5).

Schmidt, K. and C. Simone, 1996. Coordination Mechanisms: Towards a Conceptual Foundation of CSCW Systems Design. Computer Supported Cooperative Work: The Journal of Collaborative Computing 5: 155-200.

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

Star, S. L., K. Ruhleder, 1996. Steps toward an Ecology of Infrastructure: Problems of Design and Access in Large Information Systems. Information Systems Research

Star SL and GC Bowker, 2002. How to Infrastructure in The Handbook of New Media.. Lievrouw and Livingstone (eds), SAGE Publications, London, p151-162

Timmermans, S. and M.Berg, 2003. The Gold Standard: The Challenge of Evidence-Based Medicine and Standardization in Health Care. Philadelphia: Temple University Press.

UKOLN, United Kingdom Office for Library Networking, Looking at Interoperability. URL: http://www.ukoln.ac.uk/interop-focus/about/flyer-interoperability.pdf

University of Pittsburg: Dean Larsen was interviewed on the topic of the I-Schools movement at the Coalition for Networked Information (CNI) Task Force meeting on December 5, 2005. The 27 minute interview is at: http://connect.educause.edu/node/1717#attachments. The full set of interviews recorded at the meeting is available at: http://www.educause.edu/podcasts?msg=resources.

Webster's Ninth New Collegiate Dictionary, 1991. Merriam-Webster, Springfield, MA.

White, J., 2003. Task Force on Science and Engineering Infrastructure. http://www.nsf.gov/nsb/documents/2003/start.htm

Wilson, T.D., 1997. Information Management. In International Encyclopedia of Information and Library Science. London: Routledge, p. 187-196.

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

Zimmerman, A.S., forthcoming. New Knowledge from Old Data: The Role of Standards in the Sharing and Reuse of Ecological Data. Science, Technology, & HumanValues.

References Notes

1. Data work (e.g. Bowker and Star, 2000; Bowker 2001; Penniman, 2002; Parsons and Duerr, 2005), information systems (e.g. Lyytinen, 1987; Friedman, 1989; Iivari, 1991; Khazanchi and Munkvold, 2000; Choo, 2002; Chen and Hirschheim, 2004), data reuse (e.g. Birnholtz and Bietz, 2003; Zimmerman 2003, 2005), information management (e.g. Macevičiūtė and Wilson, 2005; Michener et al, 1994; NRC, 1995; Black and Brunt, 1999; Michener and Brunt, 2000), digital preservation (e.g. Granger, 2002; MacLean and Davis, 1998), information technology (e.g. Morton, 1996), and infrastructure (e.g. Hughes, 1993; Star and Bowker, 2002; Edwards, 2003)

2. Sociotechnical research, social informatics, and infrastructure studies (e.g. Kling, 2000; Karasti and Baker, 2004; Sawyer 2005; Bowker et al, in preparation)

3. Infrastructure (White, 2003) and cyberinfrastructure (Atkins, 2003, Pfirman, 2003, Futrell, 2003, David, 2004; Berman and Brady, 2005), on long-term data collection and curation (Hedstrom et al, 2002; Lord and Macdonald, 2003; NSB 2005; Karasti et al, forthcoming), on collaboratories (Finholt 2003, 2004) and e-Science (Hey and Trefethen, 2003).

4. Long-term data stewardship (Lord and MacDonald, 2003; NSB, 2005; NSF, 2003).

5. Design and standard-making processes from the fields of organization science, information systems and action research (Weick, 1995; Greenbaum and Kyng, 1991; Fujimura, 1992; Brunsson and Jaccobson, 2000; Timmermans and Berg, 2003) and from analytic ethnography (Karasti et al, 2004; Baker et al, 2005).

6. Participatory design (Schuler and Namioka, 1993; Dittrich et al, 2002), Computer Supported Cooperative Work (Kaplan and Seebeck, 2001; Schmidt and Simone, 1996), and User-Centered or Meta-Design (Fischer, 2003)

7. Appendices

7.1 Appendix: Glossary of Terms

Term	Definition
Collaborative design	a user-centered approach that draws from participatory design practices while foregrounding a participant research partner providing site feedback in addition to observer providing broad frameworks and critical perspective. Can be referred to as co-design or learning-by-design.
Data curation	concern with preservation and availability of data over the long-term; work with scientific measurements analogous to work of library science that has traditionally managed written documents and now oversees digital artifacts as well
Data heterogeneity	refers to multiple differences: among data collections and parameters within a collection, among data recorded according to varying criteria in terms of sampling, procedures and precisions; among data available resulting from a diversity of data records, descriptions, and analyses as well as from a diversity of models for storage, access and exchange.
Data interoperability	refers to data defined and stored in such a way that like types can be identified for comparison and/or combination; it entails community conventions and standards, syntactic and semantic agreements. See interoperability.
Data management	involves data handling (organizing, classifying, storing, labeling, and cross-indexing) and control (data definition, requirements, and quality) in order to make data available for use
Data stewardship	the management of data and its context, from planning to creation, organization, preservation, curation, and/or loss, involving data capture and describing, transformation and networking, use and reusing. Stewardship may center on work in the field, laboratory, project, program, center, or network but extends to concern with the dynamics and interplay of the whole system.
Design	Refers in data work to the ability to create data models, systems and infrastructures as well as to work with multiple perspectives of participants, practices, and standards. Both for individual and collaborative data efforts, design involves the theory and practice of

	identifying a purpose, planning strategy, and analyzing results while arranging parts and developing end products such as a database schema or a web interface.
Informatics	The science of gathering, classifying, manipulating, modeling, preserving, and presenting as well as designing for and teaching about data and information. It is the application of information science including information technologies and of social science including collaborative techniques to a field of study in ways that promote organization and flow of data, that highlight design of information systems and environments, and that draw upon concepts and methods to support data stewardship as well as the needs of data producers and users.
Informatics infrastructuring	adopted in our work with informatics as an active process of creating infrastructure that highlights the combined ongoing involvement needed to create and sustain a functional informatics environment.
Information infrastructure	Consists of intertwined technical and social components with organizational and conceptual aspects involving individuals and community participants carrying out work – designing, implementing, using, maintaining and redesigning – that frequently addresses the interface of human and information systems. See infrastructure and infrastructuring.
Information management	involves the use of design, management, mediation, and communication principles for development of information systems as well as coordination of data flows that take into account distributed networks and long-term approaches to data access, preservation and exchange represented via data conventions, guidelines, protocols, standards, and best practices.
Information science	a term whose definition, scope, and terminology is in development; taken here to be a domain of science incorporating computer, library, and systems science along with data, knowledge, and information management together with technological, organizational, and social configurations. Organizational elements include school districts, regional power grids, and the internet.
Information technology	application of computer science, development of technological architectures, and implementation of computer technology; includes work with equipment, firmware, software, and data transport mechanisms as well as arrangements of digital networks and their interfaces; involves design, development, and maintenance of equipment and systems focusing on automated handling of information
Infrastructure	a broad category referring to pervasive enabling

	arrangements and services. In a community arena there are schools, fire and police departments. Physical supports include pipes, wires, and roads. In a digital arena, examples are computational services and help desks as well as data and informatics resources. See Information infrastructure.
Infrastructuring	a term coined to emphasize the active nature of creating infrastructure addressing sociotechnical dimensions (data heterogeneity issues, classification paradoxes, local practices, and the making of standards and knowledge) in addition to creating traditional physical constructions, ie pipes and wires.
Interoperability	refers to a state or the goal of ensuring that the systems, procedures and culture of an organization are managed in such a way as to maximize opportunities for exchange and reuse of data and information (see UKOLN).
Mediator	the role that prompts and/or facilitates joint sense-making through articulation and translation aimed at negotiating, interfacing, and aligning flows of information; mediation occurs between data collecting, preserving, and delivery; between data types, formats, and structures; between data collectors, curators, and users; between technologies and information workers of all types
Ocean Informatics (OI)	the application of informatics to the domain of Ocean Science. It is the work that occurs at the intersection of oceanography, science studies, and information management.
User-centered design	an approach to design of data and information systems focusing on user practices, needs, and participation. This design process fosters learning through articulation and incorporates ongoing assessment. As an integral part of interactive user-centered design, there are socio-cultural-historical perspectives and collaborative techniques, prompts reflection, and enhances assessment as part of an iterative design cycle.

7.2 Appendix: Lessons 1972-2005

Toward an Understanding of Informatics:
Lessons, Principles, Features, Elements, Findings, and Strategies

I. Common Findings with Articulation

Drawn from Bateson's three levels of learning: 1-learning something; 2-learning about learning something; 3-learning about different approaches to learning about learning something (Bateson, G. Steps to an Ecology of Mind. New York: Ballantine, 1972). Multilevel tensions resulting in infrastructural transcontextual syndrome:

1. The gap between designers and users (articulation of concept that travels from level 1 technical instructions to level 2 user questions)
2. The gap inherent in discussions within community (articulation of contextual translations from levels 1 to 2 to 3)
3. The gap between routines and rapidly-growing infrastructures (articulation of change)
4. Double levels of language in design and use (articulation of ambiguity)

II. Ten Principles of Sociotechnical Design (Cherns, 1987; 1976)

1. Compatibility
   The process of design must be compatible with its objectives. This means that if the aim is to create democratic work structures design must be participative.
2. Minimal Critical Specification
   Specify only what is essential.
3. Variance Control
   Variations from norms should be handled close to the origin of the data.
4. Boundary Location
   Boundaries used for definition of element should not impede the flow of data, knowledge, or information.
5. Information Flow
   A design aim is to facilitate the flow of data and information.
6. Power and Authority
   Information must go to those who need to take action.
7. Multifunctionality
   Organizational adaptability is enhanced by participant roles that are flexibile and respond to change through new skills.
8. Support Congruence
   Keep I mind human values by providing opportunities to learn and rewarding the insight and understanding of team participants rather than solely what is at hand.
9. Transitional Organization
   Plans must include consideration of transition to future systems with change viewed as a design opportunity or opportunity to reflect upon and improve existing designs.
10. Incompletion

(Re)design should be recursive, with regular review and evaluation.

III. Eight Salient Features of Infrastructure (Star & Ruhleder, 1996; Star & Bowker, 2002)

1. Embeddedness
2. Transparency
3. Reach or scope
4. Learned as part of membership
5. Links with conventions of practice
6. Embodiment of standards
7. Built on an installed base
8. Becomes visible upon breakdown

IV. Framework of Elements for Information System Analysis (Iivarii, 1991; Karasti, 1994; Baker and Bowker, in press)

1. Methodology
    - Constructive methods: conceptual dev, technical dev, triangulation
    - Nomothetic methods: formal-mathematical, experiments, field studies/surveys
    - Idiographic methods: case studies, action research
2. Ethics
    - Role of IS science: means-end oriented, interpretive, critical
    - Values of IS Research: org/mgmt oriented, user oriented, others (educative)
3. Epistemology
    - Positivism
    - Anti-positivism
    - Multiperspectivism
4. Ontology
    - View of information/data: descriptive facts, constitutive meanings
    - View of information/data system: technical system, organization/social system
    - View of human beings: determinism, voluntarism
    - View of technology: technological determinism, human choice
    - View of organization and society: realism, structuralism, interactionism

V. Seventeen Lessons from Open Source (Raymond, 1999)

1. Every good work of software starts by scratching a developer's personal itch.
2. Good programmers know what to write. Great ones know what to rewrite (and reuse).
3. "Plan to throw one away; you will, anyhow." (Fred Brooks, The Mythical Man-Month, Chapter 11)
4. If you have the right attitude, interesting problems will find you.
5. When you lose interest in a program, your last duty to it is to hand it off to a competent successor.
6. Treating your users as co-developers is your least-hassle route to rapid code improvement and effective debugging.
7. Release early. Release often. And listen to your customers.

8. Given a large enough beta-tester and co-developer base, almost every problem will be characterized quickly and the fix obvious to someone.
9. Smart data structures and dumb code works a lot better than the other way around.
10. If you treat your beta-testers as if they're your most valuable resource, they will respond by becoming your most valuable resource.
11. The next best thing to having good ideas is recognizing good ideas from your users. Sometimes the latter is better.
12. Often, the most striking and innovative solutions come from realizing that your concept of the problem was wrong.
13. "Perfection (in design) is achieved not when there is nothing more to add, but rather when there is nothing more to take away."
14. Any tool should be useful in the expected way, but a truly great tool lends itself to uses you never expected.
15. When writing gateway software of any kind, take pains to disturb the data stream as little as possible - and *never* throw away information unless the recipient forces you to!
16. When your language is nowhere near Turing-complete, syntactic sugar can be your friend.
17. A security system is only as secure as its secret. Beware of pseudo-secrets.

VI. Department of Navy Scope of Data Challenges Lessons Learned (DON, 1999)

1. Data problems are not unique to any one functional area or organization.
2. There is a need for policy, process, supporting infrastructure, and a plan to leverage efforts.
3. Data management requires senior management champions.
4. Data management is not adequately addressed in budget or acquisition processes.
5. In an era of network-centric warfare, addressing the issues has never been more essential.

VII. Sociotechnical principles for system design (Clegg, 2000)

Meta-principles

1. Design is systemic.
2. Values and mindsets are central to design.
3. Design involves making choices.
4. Design should reflect the needs of the business, its users and their managers.
5. Design is an extended social process.
6. Design is socially shaped.
7. Design is contingent.

Content principles

8. Core processes should be integrated.
9. Design entails multiple task allocations between and amongst humans and machines.
10. System components should be congruent.
11. Systems should be simple in design and make problems visible.
12. Problems should be controlled at source.
13. The means of undertaking tasks should be flexibly specified.

Process principles

14. Design practice is itself a sociotechnical system.
15. Systems and their design should be owned by their managers and users.
16. Evaluation is an essential aspect of design.
17. Design involves multidisciplinary education.
18. Resources and support are required for design.
19. System design involves political processes.

VIII. The Science of Information Management Recommendations (Graves et al, 2002)

1. Knowledge representation and management
1.1. Recommendation: Develop interoperable identifier and metadata systems.
1.2. Recommendation: Code and preserve institutional and personal knowledge about data in order to support effective secondary and long-term archival usage.
1.3. Recommendation: Create methods for the management of data generated by new forms of instrumentation such as MEMS (micro-electromechanical systems).

2. Data integration
2.1. Recommendation: Automate the rapid integration of new sources and services.
2.2. Recommendation: Develop self-integrating software systems.
2.3. Recommendation: Develop techniques for the composition of services.
2.4. Recommendation: Support the integration of heterogeneous source/services through a spectrum of approaches spanning standards, integration technologies and interoperability techniques.
2.5. Recommendation: Develop application specific intersection articulation among data sources.

3. Collaboration and security
3.1. Recommendation: Shift to a collaboration-centric world, in support of human collaboration for specific objectives/tasks/goals.
3.2. Recommendation: Push the state-of-the-art of information security and trust.

4. Usability
4.1. Recommendation: Understand the importance/relevance to the user through advanced techniques.
4.2. Recommendation: Enhance decision-making tools with the ability to make predictions and support multiple hypotheses.
4.3. Recommendation: Create an environment for conducting large-scale experiments that involve organizational and process elements of technologies.

5. Application-oriented research
5.1. Recommendation: Bring to each application area, the strengths of the other two.
5.2. Recommendation: Fund a small number of large-scale experiments across application areas.

IX. Directions for Digital Libraries: Basic Themes in Long Term Research (NSF Digital Libraries, 2003)

1. Understanding information and its uses
2. Appropriate stewardship over information
3. Fitting technology-enabled opportunities into the social fabric
4. Matching system capabilities to user needs
5. Interoperability

X. Social Informatics Common Findings (Sawyer, 2005)

1. The phenomenon of interest will vary by the level of analysis. (5)
2. The differential effects of the design, implementation and uses of ICTs often have moral and ethical consequences. (3)
3. The design, implementation and uses of ICTs have reciprocal relationships with the larger social context. (4)
4. Uses of ICT lead to multiple and sometimes paradoxical effects. (1)
5. Uses of ICT shape thought and action in ways that benefit some groups more than others. (2)

XI. Informatics Good Practices (Baker, 2005, this report)

1. Incorporate data problem formulation and data scoping early in the scientific planning process.
2. Recognize articulation, translation, negotiation and mediation as central to work with data.
3. Partner with appropriate information professionals for data work.
4. Create collaboration opportunities as well as coordination mechanisms for community work.
5. Recognize informatics as conducting research while carrying out information management.

7.3 Appendix: Emergent Academic Informatics Arenas

Informatics and Related Disciplines

Information is a confounding entity-concept that defies categories and crosses boundaries. Topics pertinent to informatics and information infrastructure are boundary-crossing; issues such as networks, systems development, communication, and collaboration along with data practices, curation, integration, exchange, and synthesis overflow from the confines of libraries, computing centers, and disciplines. Information science research in general and informatics in particular are needed to pose new questions as well as to model data and to prototype solutions. To meet these needs, there is an emergence within academic arenas of schools of information, information studies departments, and informatics programs - standing alone or in conjunction with library science, engineering, and/or computer science departments.

Despite a well-established presence, the iniquitousness of data, knowledge, and information makes difficult the instinctive human approach to organization using categories ordered by hierarchical relationships. Cross-category concepts leave us pondering whether computer science is a subset of information science or of engineering, whether library science is to become 'library and information science' or whether it is to become 'knowledge management', and just where to place the related work of information technology and of information systems.

Institutional approaches to study of information have taken a variety of forms over the last decade. Early information work occurred in Departments of Library Science, Engineering, Computer Science and Geography Departments but is beginning to appear in new guises. Within Computer Science there are Departments of Computer and Information Science (e.g. University of Pennsylvania; University of Oregon) and within Schools of Library Science there are Departments of Information Studies (e.g. UCLA). Traditionally separate, domains are being merged as with the School of Library and Information Science (e.g. Indiana University), the Information and Library Science School (e.g. University of North Carolina) and Schools of Information and Technology with Schools of Science and Technology Studies (e.g. Virginia Tech) and with Business Schools with Departments of Information Technology Management (e.g. University of Hawaii). There are Schools of Informatics (University of Buffalo) and Information and Computer Science with Departments of Informatics (e.g. UCIrvine) as well as Schools of Information and Communication. There are Science Studies Programs (e.g. UCSD), Centers for Social Informatics (e.g. Indiana University), Centers for Earth System Science (e.g. UCSB), and Departments of Information Management and Systems (e.g. UCBerkeley).