20-22May 2002, Proceedings of the Second National Conference on Digital Government Research, 141-144pp

# Designing an Infrastructure for Heterogeneity in Ecosystem Data, Collaborators and Organizations

Karen S. Baker

Scripps Institution of Oceanography
University of California, San Diego
La Jolla, California 92093-0218 USA
kbaker@ucsd.edu

Geoffrey C. Bowker and Helena Karasti

Department of Communication
University of California, San Diego
La Jolla, California 92093-0503 USA
bowker@ucsd.edu, hkarasti@ucsd.edu

## 1. Abstract

To develop robust datasets for long-term re-use, new approaches are needed that incorporate relevant facets of organizational culture in their description. Early ethnographic research points to the importance of holding narrative accounts of data use alongside formal metadata structures. We describe our proposal to identify models for the design of information protocols and procedures within the Long-Term Ecological Research community that take account of the working practices of all the participants involved in the varied aspects of information processing.*

## 2. Introduction

Biodiversity and ecosystems data are currently being gathered in a large range of formats by a constellation of loosely connected private, government and not-for-profit agencies. The normal response to this double heterogeneity has been the development and enforcement of metadata (data about data) standards; in this response one tries to abstract data away from its organizational context in order to render it universally accessible. This project takes the opposite tack, and seeks new ways of grounding environmental data in its organizational context in such a way that it can both be used more flexibly today and so it can retain value longer. The hypothesis, based on the last 25 years of work in the field of Science Studies, is that formal data descriptions must be 'wrapped' in informal descriptions in order to be useful. The goal of this project is to open up the database inquiry of the biodiversity and ecosystems communities generated by their need for very long lasting and highly distributed data. We focus on communications in ecosystem informatics through the use of structural (e.g. standardized classifications; metadata) and alternative (e.g. narrative) methods. Our approach is action-oriented research that integrates ethnographic fieldwork and participatory design (Karasti, 2001). Through our theoretical interests in information ecologies and work practices, we intend to articulate connections between organizational and scientific data.

## 3. Research Approach

The issues involved in biodiversity and ecoinformatics are complex and large-scale. A recent call for setting priorities for new interdisciplinary environmental research programs points out the need for action outside the status quo of disciplinary science (Kinzig et al, 2000). We have gathered a team of investigators that shares the recognition that contemporary research questions require new interdisciplinary approaches.

Odum (1996) writes "because ecology is an integrative science, it has tremendous potential to provide a communication bridge between science and society". We extend his observation to include the bridging between ecological field sciences,

* NSF Grants EIA-01-31958, DBI-01-11544 and OPP-96-32763 support this work.

information sciences and social sciences. We are working at the intersection of these three sciences taking into account the distinct Communities of Practice (CoP, Lave and Wenger, 1990) that have developed at the intersections (Figure 1). Active boundary communities include those interfacing Environmental and Social Sciences (e.g. Computer Mediated Communication, CMC), Environmental and Information Sciences (e.g. Information Management, IM), and Information and Social Sciences (e.g. Human Computer Interaction, HCI; Computer Supported Cooperative Work, CSCW; Social Informatics, SI; and Participatory Design, PD) with new insights into data handling, work practice and infrastructure effectiveness. The Center (U) designates a union of understanding.

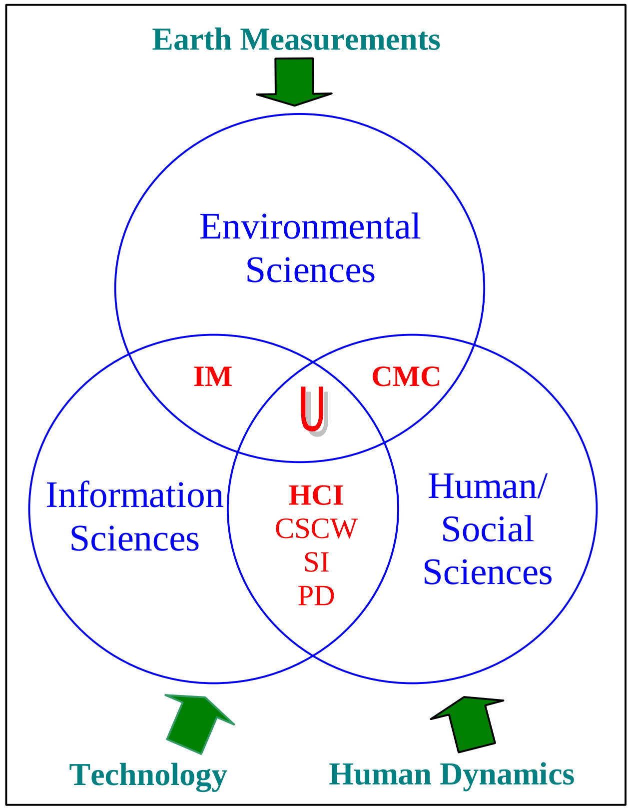


Figure 1: Science Domains and Communities of Practice (CoP)

### 3.1 At the Interface of Information Sciences and Social Sciences

The number of CoPs shown in Figure 1 at the join between Information Sciences and Social Sciences is an indicator of the complexity that is socially generated (PCAST, 1998). Social Informatics refers To the body of research and study that examines social aspects of computerization, including the roles of information technology in social and organizational change and the ways that the social organization of information technologies are influenced by social forces and practices (http://www.slis.indiana.edu/si/).

Computer Supported Cooperative Work is a term to describe the understanding of the way people work in groups with the enabling technologies of computer networking, and associated hardware, software, services and techniques. It is concerned with designing shared information spaces and supporting heterogeneous, open information environments that integrate existing single-user applications (http://www.telekooperation.de/cscw/cscw.html). Participatory Design is an approach to the design and development of technological and organizational systems that places a premium on the active involvement of workplace practitioners in design and decision-making processes (http://www.cpsr.org/program/workplace/PD.html)

Our work represents one approach to examining the issues of managing data for long-term, and eventually very long-term, use within a widely distributed, loosely connected network of scientists: the Long-Term Ecological Research (LTER) network (Franklin, 1990). We have argued (Bowker, 2001; Baker et al, 2000) that full attention must be given to the social and organizational dimensions of ecosystem informatics if we are to develop robust, reliable and useful databases for the future. Our research effort is grounded on ethnographic fieldwork. Ethnography is an approach for developing understandings of the everyday activities of particular communities of people through participant observation and interviews. The LTER is a complex working environment given the diversity of twenty-four sites, a Network Office coordinating the sites, and a range of associated partnerships. Embedded within each site and the Network Office is an active information management component. We are studying the data ecologies and work practices within the LTER, eliciting and articulating significant elements of collaboration and community, and considering design artifacts that enhance network science in order to better understand and to plan for the management of scientific heterogeneity.

How can the goal of creating and preserving meaningful long-term data be initiated by and grounded in everyday practice? We aim to articulate the relationships between data held in the computer and transmitted over a network, data held in the human mind and shared through stories as well as data held on paper and stored in file cabinets. A workshop format is planned with members of the LTER community in order to elicit multiple voices and to promote community discussion. We see workshops as an important enabling mechanism for reflection and change.

### 3.2 Research Focus

There is a wide gap in the field of ecosystem informatics between what is being produced by information technology specialists and what is actually useful to working scientists. The difficulties inherent in bridging this divide in computer science and field science partnerships are documented in retrospectives (e.g. Stonebraker, 1994) but are rarely addressed directly in practice. An interdisciplinary team working with an established, ongoing network provides the unique opportunity to focus on communications in ecosystem informatics.

The flow of information in field sciences consists of multiple steps starting with project design followed by fieldwork, moving data through ordering filters or frameworks into a digital record that may then be put through an output filter for retrieval (figure 4, Baker et al, 2000). In most scientific databases, organizational data falls away and is lost very quickly. Bowser (1986) for example, discusses problems with interpreting data predating the LTER site in Wisconsin. Measurements of lake water acidity would be different depending on whether they were taken in the laboratory on return from the field or in the field – loss of $CO_2$ in samples over a few hours changes the measurements. This information was nowhere mentioned in published reports, but fortunately Bowser and colleagues were able to locate a retired limnologist who remembered the procedure. The point here is that knowledge about the *practice* of old limnologists was needed. No-one at that time would have thought to retain this information about the data that they were collecting – everyone using the data at the time would know how lake water was collected. However, this vital information was lost over time.

Clearly, metadata standards alone will not solve this kind of problem. There is an initial awareness of this within the LTER information managers, as one of them stated: “We are finding now that the structured [metadata] is much more useful in terms of producing machine readable information but the narrative often times contains more information.” In this study we are starting to explore today’s work procedures while considering those that will be vital to scientists fifty or one hundred years from now. We are contemplating ways of preserving organizational data (defined as data about synthesis, work practice and institutional framework) without overburdening the already stretched resources and time of research projects and data management.

We view databases as communication tools for sharing data. There are two categories of sharing: the here-and-now of data collection in support of ongoing ecological research as well as the future of data re-use in answering different, as yet unasked, questions. One information manager cognizant of these two aspects articulated the following: “if people feed us back information about a dataset … we put it into the database, then other people can read what other people have said about that dataset… some of these people in the past have given us really comprehensive reviews of the data, it’s like wow, this should be part of the data, I did not know that. I’ve not had time to analyze it, so if someone takes the time to analyze it, especially an outside person, a PI might tend to do some corrections or what ever, but someone outside really sees it objectively: this does not match, this does not make sense… Another thing … the data manager knows a lot about what really are the good and bad aspects of the data. … because we have handled it, we know what works and what does not… That should be part of the metadata. Because ultimately if you don’t write those things down, they are going to get lost. … It’s stuff that is more valuable than a lot of this other descriptive information about a dataset. I mean in terms of a real quality ‘gut feeling’ of how good it is. You know, like a ‘subjective quality indicator’ of some sort.”

Recognizing the incremental change processes inherent to long-term datasets, one information manager describes: “When we now are moving our datasets into our new system, we need to go back and see the abstracts written 20 or 30 years ago …

we are not asking the same questions anymore. We need to keep the old ones and start writing more descriptive information because the thinking changes… People's thoughts on why it is being collected and should continue to be collected, change - different research questions are being asked."

This project starts to explore new ways of grounding environmental data in its organizational context so that it can both be used more flexibly today and so it can retain its value longer. It will develop into a larger follow-on study of the articulation between metadata and narrative modes of data and possible ways of representing them. This work facilitates a timely dialogue focused on "data ecology" (the relationship between data and their multiple environments) and builds toward the concept of an "organizational ecology" (the relations between data, participants and their networks).

### 3.3 Initial Findings

While we are still in the data collection cycle of our project, some initial themes are emerging from interviews and observations of work practice. The formal work practices of LTER information managers relate to gathering, quality analysis and quality control, archiving and facilitating data exchange. In themselves, these comprise a demanding set of tasks. However, even more demanding are the inherent, continuing tensions between the formal and the informal work of creating and holding the organizational memory encompassing local datasets and the network information in general. This suggests the need to find ways to characterize and represent informal work that would enable long-term data use. Within the Library and Information Science community is a growing "awareness of the immense scope of the potential preservation crisis" (http://www.clir.org/ pubs89/contents.html) with 'incremental metadata considered a key to successful migration of data. Such incremental metadata will take many forms in the attempt to determine the appropriate mix of structured and narrative accounts of datasets.

## 4. Conclusion

We propose to identify pertinent types of contextual information relevant to data synthesis and promote community discussion to enhance representation of dynamic, multi-level aspects of ecological data. As is the case with action research, a dual level approach is planned: both the practical level as represented by our work of learning with and observing the LTER as well as the conceptual level which transcends scientific community. Such an approach sets the stage for asking whether there are methods complementary to logic-based approaches to information retrieval that can encompass contextual understanding, including both lived experiences and historical understandings.

Acknowledgement: We offer special thanks to the LTER community members for contributing their time and insight.

## 5. References

Baker, K.S., B.J. Benson, D.L. Henshaw, D. Blodgett, J.H. Porter, and S.G. Stafford, 2000. Evolution of a multisite network information system: the LTER information management paradigm, Bioscience, 50(11), 963-978.

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

Bowser, C., 1986. Historic Data Sets: Lessons from the past, lessons for the future, Research data management in the ecological sciences W. Michener (ed.) University of South Carolina Press, The Belle W. Baruch library in marine science 16: 155-179.

Franklin, J.F., C.S. Bledsoe, J.T. Callahan, 1990. Contributions of the long-term ecological research program - an expanded network of scientists, sites, and programs can provide crucial comparative analyses, Bioscience 40(7): 509-523.

Karasti, H., 2001. Increasing sensitivity towards everyday work practice in system design. PhD thesis, http://herkules.oulu.fi/isbn9514259556.

Kinzig, A.P., S. Carpenter, M. Dove, G. Heal, S. Levin, J. Lubchenco, S.H. Schneider, D. Starrett, 2000. Nature and Society: An Imperative for Integrated Environmental Research, Workshop Report, http://lsweb.la.asu.edu/akinzig/report.htm.

Lave, J. and E. Wenger (1990). Situated Learning: Legitimate Peripheral Participation. Palo Alto, CA, Institute for Research Learning: 1-38.

Odum, E., 1996. Ecology: A Bridge Between Science and Society. Sinauer Associates, MA.

PCAST, 1998. Teaming with Life: Investing in Science to Understand and Use America's Living Capital. President's Comm of Advisors, Science and Technology, Panel on Biodiversity and Ecosystems.

Stonebraker, M., 1994. Sequoia 2000 -- A reflection on the first three years, Seventh International Working Conference on Scientific and Statistical Database Management, 28-30 September '94, Charlottesville, VA, pp 108-116. IEEE Computer Society Press, Los Alamitos.