Through the Looking Glass: What do we see, What have we learned, What can we share? Information Management at the Shortgrass Steppe Long Term Ecological Research Site

Susan G. Stafford, Nicole E. Kaplan, and Christopher W. Bennett
Department of Forest Sciences, Colorado State University, Fort Collins, Colorado 80523-1470

Abstract

This paper documents the development of a successful information management system at a Long Term Ecological Research (LTER) site that has a rich history of data collection and management. Over sixty years of data from three separate projects are incorporated into the Shortgrass Steppe (SGS) LTER information management system and databases. People with different strengths and expertise ranging from clerical administrator, programmer, to ecologist, have filled the role of Information Manager (IM) at the SGS-LTER. Today the information management needs of the SGS are provided by a team of IMs with various levels of expertise in a wide variety of domains from information technology administration to education and outreach. It is critical for IMs at any long-term research site to understand how information and data were managed in the past and what recent changes have been added to the system, in order to effectively implement a management plan for the future. We are able to evaluate the effectiveness of different approaches to information management and have a commitment to share our successes with the information management community.

Keywords: Information management, data sets, metadata, databases, long term ecological research, grasslands

Introduction

A functional information management system must be well organized, yet nimble enough to support ecological research efforts that change with advances in technology (Stafford 1986a, b). Data collection methods at the SGS-LTER have evolved from paper field forms and notebooks to automated data loggers and downloadable satellite images. As a result of technological advances in scientific equipment, the stream of incoming information to manage and data to archive has grown in volume and complexity. The community of IMs across the LTER Network and other organizations is able to share what they have learned, as well as continue to develop tools to better manage and synthesize more data and information. As a site, we are able to look back to see how research projects, data collection and management have changed over sixty years.

History of research programs and issues

Since 1939, research has been conducted on the Central Plains Experimental Range (CPER), a 15,500 acre site located in the northeast corner of the Pawnee National Grasslands, 13 kilometers northeast of Nunn, Colorado. The CPER research site was established in response to the impact of drought and overgrazing across the Great Plains during the era of the Dust Bowl (Figure 1). Researchers from the United States Forest Service studied ways to improve management practices and promote rangeland sustainability across the fragile landscape. Early research projects focused on understanding native plants, recovery on abandoned plowed fields, and techniques for measuring plants' responses to grazing by cattle. Today, cattle from the local grazing association still graze the research site and grazing is considered a "treatment" for some long-term experiments.

In 1968, the CPER began working with the Colorado State University's Natural Resources Ecology Lab (NREL) on studies funded by the National Science Foundation including the International Biological Program (IBP) and later the LTER program (Figure 1). During the IBP, research projects were started to gain an understanding of grassland ecosystems. Over 70 scientists collaborated to obtain information from field studies that examined ecosystem interactions and grassland productivity. The IBP pioneered the ecosystem analysis approach, which brought together scientists from diverse disciplines and organizational structures (Van Dyne memo 1970). Some of the goals of IBP focused on understanding the productivity and the economic importance of grassland ecosystems. Grasslands such as the CPER and Pawnee National Grasslands, were studied within the context of human use and resource management.

The LTER program has less of a utilitarian approach to understanding and studying the shortgrass steppe compared to the IBP (personal communication, W.K. Lauenroth). LTER scientists focus on five core areas of research and monitoring that are mandated by the National Science Foundation for the LTER program. Callahan (1984) organized these foci as patterns and controls on primary production, organic matter accumulation, inorganic inputs and transport, disturbances, and populations. Since 1982, research projects at the SGS-LTER have been designed to answer questions about how factors that regulate structure and function and coupling of biotic and abiotic components in the shortgrass steppe ecosystem vary over space and time, and determine vulnerability to changes (Figure 1). Field data are collected from long-term plots on the SGS-LTER, across regional transects spanning the Great Plains, and in satellite images.

As a network, LTER researchers seek and conduct long-term continuous measurements and analyses of ecological patterns and processes at different spatial and temporal scales. They also emphasize the integration and synthesis of results within and among specific sites and the generalization of results over broader spatial scales and for a broader audience. Together IMs and researchers in the Network effectively formulate questions, coordinate data collection, manage and access electronic archival information, and exchange complex data sets.

SGS-LTER IMs maintain data sets from a variety of

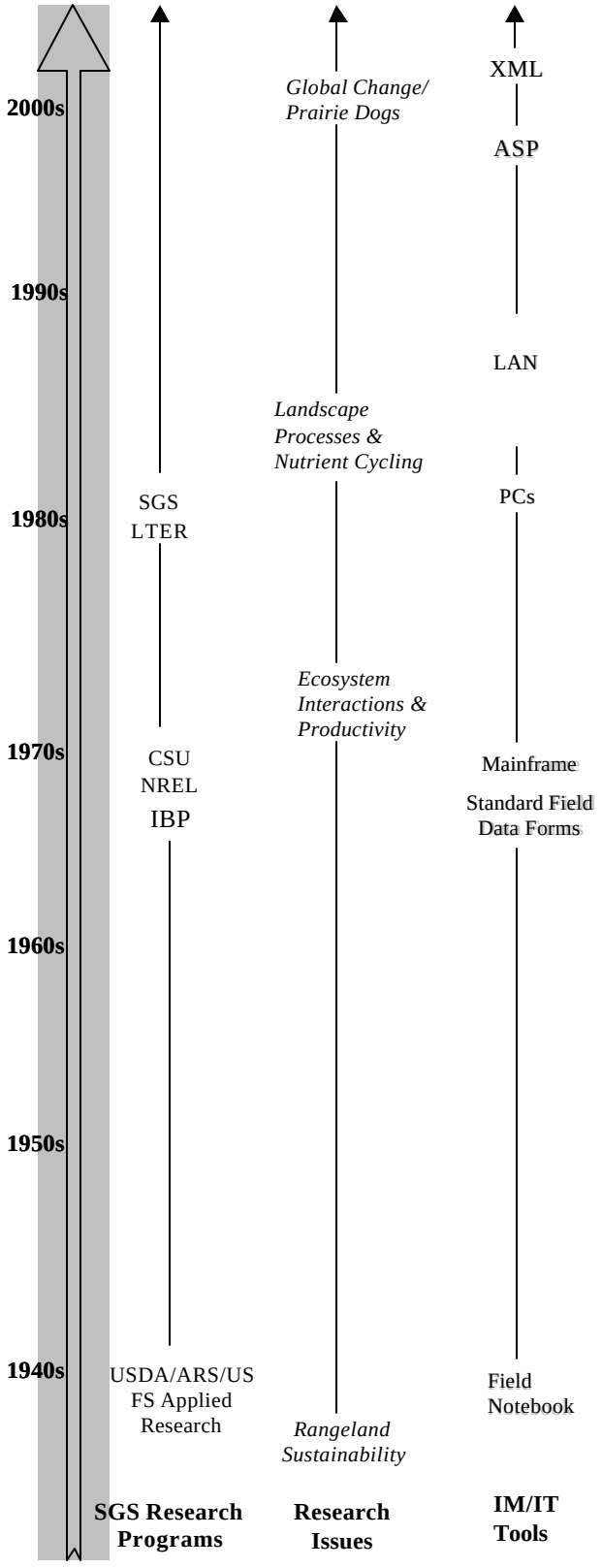

Figure 1. Development of research programs, research issues, and imformation management and technology tools from 1939 to present.

studies dating back to 1939 (Figure 2). Over one hundred short-term and long-term experiments have been or are being performed. Data sets collected earlier in time helped CPER researchers gain an understanding of what drives production in the SGS ecosystem. Additional data sets were collected since the 1960s during the IBP to gain a more complete understanding of the ecosystem. The LTER now manages many legacy data sets, and data sets generated by cross-site efforts between multiple LTER as well as other community sites. Research projects such as net primary production or meteorological monitoring are conducted at all LTER sites simultaneously and produce data and metadata that are synthesized and published for the entire LTER Network and greater scientific community (Henshaw et al. 1998). SGS-LTER IMs give high priority to working with researchers and IMs from other sites to manage information in a way that will organize and maintain clear and accurate data and metadata for use in synthesis projects, publications, and by the broader ecological community in the future.

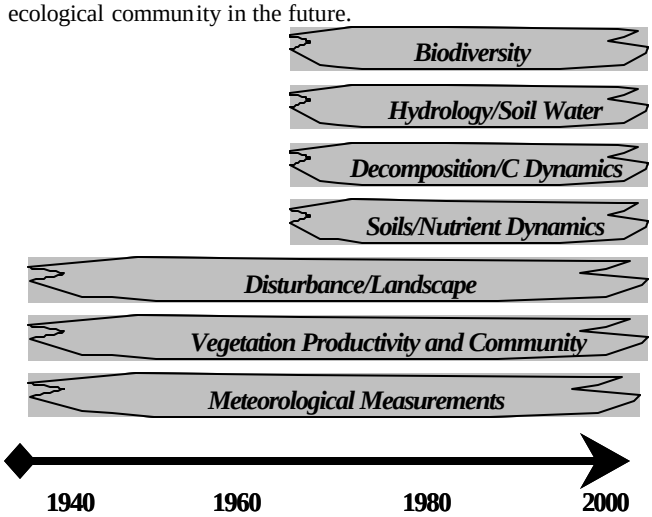

Figure 2. Data sets have been collected from research conducted on the CPER, and during the IBP and LTER programs. Diverse categories of data sets continue to be collected.

History of Information Management and Technology From CPER to LTER

The SGS-LTER information management system serves a community of educators, students, researchers, land managers, and public policy makers and manages data and other information from over sixty years of research. Data from early studies conducted during the initiation of the CPER were collected in field notebooks and on paper. Much of the data still remain in original format and research methods, results, and other information have been published by the United States Department of Agriculture and Agricultural Research Service (ARS) in technical reports.

An infrastructure for the IBP was available through facilities at the new Natural Resource Ecology Laboratory (NREL) at Colorado State University (CSU) and at the time,

state-of-the-art IBP buildings at the CPER site. The NREL clerical office staff supported researchers' efforts to integrate and synthesize data and information from seven satellite field sites in different grassland regions of North America (Pacific Northwest Bunchgrass, High Mountain Grassland, Mixed Prairie, Desert Grassland, Tallgrass Prairie, Shortgrass Prairie, and Annual Grasslands). Initiated in the planning phase of the IBP project, data management, publication, and chemical and statistical analyses were identified as needed along with support services and administration of the project. Two years into the project, it was reported by Van Dyne (1970) that researchers recognized, "…more and more opportunity to undertake integrative synthetic activities which require utilizing data from different trophic levels, from different sites, or combinations of these." IBP researchers then created a policy to share data and complete cooperative analysis and review (Van Dyne 1971). IBP scientists were expected to share data, keep data accessible to be used again, and disseminate information through publishing a series of technical reports and peer reviewed articles.

However, there were difficulties. In 1970, Van Dyne and the IBP collaborators realized the new challenges of working with a large group of scientists from various disciplines. Project information and data needed to be managed within an organizational framework designed to produce broader knowledge and optimize resource management in various grassland ecosystems. Problems arose while meeting the challenge of managing all of the data that IBP produced. As IBP continued, a significant amount of data and information were generated, synthesized, and then published in various reports and publications and archived on Fortran card decks. This system made it very difficult to locate information. In 1972 a keyword index was generated and one volume of abstracts was produced in an effort to collate and categorize the information in one place. It also was difficult to get senior researchers to endorse data sharing, to maintain field equipment, and to invest in adequate support staff to meet the needs of the project. To solve these problems, program directors sought to keep researchers stimulated and dedicated to the project by providing more services to researchers, including data and sample analyses. Involvement in education and outreach activities at the university and resource management agencies, was possible because of site support, as well as the central storage of information.

At the time, available technology limited storage of IBP data and information to hard-copies of field data collections forms, Fortran card decks used for analysis on a mainframe computer at NREL, an extensive series of technical reports, and a closet full of over two dozen 7-track tapes. In total, more than 80 data formats, 111 theses and dissertations and over 200 publications were generated. The IBP project generated such large amounts of data and information between 1969 and 1978 that the SGS-LTER was able to use background information to plan new research strategies (Chaffee 1984).

In 1980, the National Science Foundation selected six sites to establish a Network of LTER sites. The SGS was in the first cohort of LTER sites. From the onset of the LTER Network (Callahan 1984), data management and accessibility were identified as major components of the LTER mission with the LTER IM mission explicitly stating "our goal is to promote ecological science by fostering the synergy of information systems and scientific research." (Baker et al. 2000). The first major undertaking of SGS-LTER information management was to finish documenting and indexing the hoards of IBP data that went un-catalogued after funding ran out. The IBP 7-track tapes were converted to 9-track tapes and more current data were stored on-line on a 10-megabyte removable disk pack that was indexed by project and study site. Metadata were stored where possible. Data managers also recognized that it was almost impossible for secondary users to navigate this data management system to location information of interest (Chaffee 1984). They looked to the future, when user-friendly, expandable, and visible tools would be developed.

By the late 1980s, the SGS-LTER sought to develop a truly integrated computer environment using a LAN (Local Area Network). Researchers built a regional model and a site-level Geographic Information System (GIS) database with data they collected over different spatial and temporal scales. There was an interest in applying new computer technology to ecological research through simulation modeling, spatial analysis using GIS, and remote sensing (Burke 1992, Kirchner 1989). SGS-LTER acquired the hardware and software to build a computing network to support these efforts and began to think about managing data and information from site, regional, and network-wide levels. Computer programmers developed software for models (e.g. Century) and analysis that could be transferred easily to other types of computers such as the IBM PC and Sun Workstation, rather than a mainframe.

IMs and researchers sought to publish data and information by means of hosting field trips to interest groups; promoting newspaper, television, and radio coverage of their research efforts; and publishing in the CSU magazine. Scientific and technical information was disseminated throughout the ecological community through synthesis activities, publishing in scientific journals, and working with an interdisciplinary group of researchers on simulation models and regional analysis.

The 1990s brought a new decade of technological advances, ecological issues, and synthesis research. SGS-LTER researchers and scientists from other LTER sites were generating and manipulating more complex data in models that predicted environmental factors from soil water movement to the impact of global climate change (2002-2008 SGS-LTER Proposal). The abilities of the SGS-LTER information management system were enhanced by hardware and software capable of mass storage, management of spatial and point data sets, increased computing power and speed, and the digitizer, which created digital information on the fly.

In the earlier days of the SGS-LTER computer network, the dominant work platform consisted of a network of Sun unix workstations. While Personal Computers (PCs) played an increasing role over time for each individuals' work, as of early 1999 the SGS-LTER unix file services network had grown to about 20 separate physical drives that were attached to many different physical computers and networked into a model via NFS

("network file system"--the networking part of the unix workstation operating system). SGS-LTER expanded their storage capacity by buying individual drives as money became available, without any deliberate planning to centralize or organize the storage of information. Often, space for projects would be allocated on the "homestead" model, i.e., the drive with the largest available free space would be used for the next new project. PCs added complexity to the data management system by further decentralizing the location of data. This resulted in a network of 20 to 30 drives used for managing various data sets and any metadata associated with SGS-LTER projects. Another problem, consistent throughout each research program through time, was the lack of adequate project documentation, i.e. metadata. Some metadata were recorded for GIS data sets, but as an operating principle the process was essentially lacking, making it difficult to re-use old data. This has been recognized as contrary to the SGS-LTER mission, and so it has been addressed in the new data storage model.

We are able to meet the current and future needs of SGS-LTER well into the foreseeable future with a new storage technology called RAID (Redundant Array of Inexpensive Disks). The product is a single box that contains 5 or more large-capacity commodity disk drives made to operate as a single physical drive. The new data storage model organizes the total storage system into hierarchical components by identifying key attributes for categorization. In addition, some physical optimization is gained by defining an expected pattern of frequent versus infrequent file modification and access. The system is organized in a manner that allows flexibility for developmental data, while recognizing the need for growth and stability of finished data products.

After data are entered and processed through a single point of entry on our new data storage network and assured for quality, the information is transferred to the SGS-LTER Microsoft Access RDBMS (relational database management system) and may be downloaded from our web site (sgs.cnr.colostate.edu). Corresponding attribute definition tables and metadata are included in the database and also are accessible through our website to download in ASCII or XML format.

History of the people from CPER to LTER

In 1939, researchers at the CPER site did not think about archiving, sharing, or synthesizing their data sets or metadata with other scientists or institutions. The ARS who after 1953 administered the CPER, keeps most of the original data sheets in safe storage. Staff at CPER and SGS-LTER are currently transferring CPER data from paper to electronic format. Administrators at IBP realized a need for a data management policies. They organized published information in indexes and attempted to use whatever limited technology was available to convert information from data sheets to a digital format. Programmers during the first SGS-LTER project (LTER I, 1982-1986) continued efforts to convert information from IBP to a more stable format.

In the beginning of LTER I, the SGS-LTER data management system focused on three issues: (1) reduce software development costs by creating utilities that can serve several types of data; (2) centralize processing and screening of data in a timely manner; and (3) safeguard data documentation to avoid loss or corruption. By LTER II (1986-1990), researchers and IMs recognized the need to archive and communicate research information from project initiation. Support staff became more involved with researchers during data and sample collection, verification, lab processing or data entry, QA/QC, archival, and publication (Brunt 2000). SGS-LTER researchers set goals and policies for the IM personnel. Researchers and IMs formed teams to assure data quality, documentation, and timeliness. LTER II began to use data collected in the field to develop models. Programmers again served as IMs and helped researchers to synthesize data and develop models.

After the first two funding cycles, LTER III (1990-1996), IV (1996-2002), and V (2002-2008) placed new demands on SGS-LTER IMs (Figure 1). Our IM system has evolved since the late 1980s, when we saw major technological advances such as faster and larger capacity PCs and the internet. Through the early 1990s, during LTER III, IM personnel consisted of two programmers with backgrounds in biological sciences. Just prior to LTER IV, IM staff became involved in the field and laboratories, and at all steps from project initiation to publication and archival information (Figure 3).

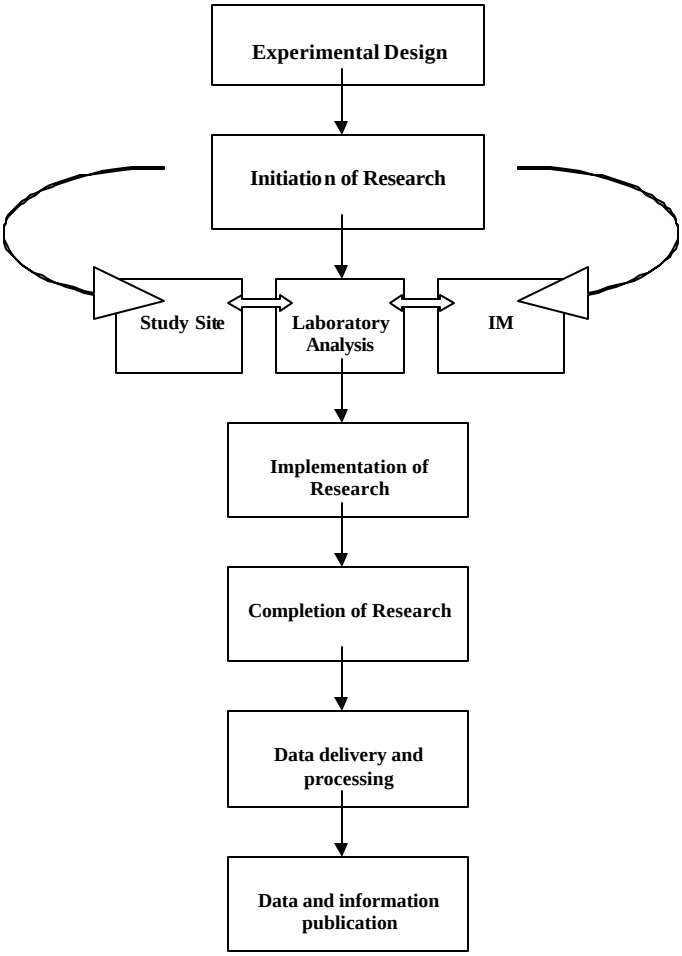

Figure 3. Systematic approach to involvement of staff, information managers and researchers in all steps from project design to data and information publication.

As technological tools developed, our approach changed. Originally, we needed programmers because there were no off-the-shelf products to meet our needs. More recently, the focus of the IM personnel changed to the development and construction of hierarchical relational databases for legacy data from CPER and IBP and new data sets collected throughout the Great Plains region. The development of the World Wide Web created a new venue to publish data and empowered every LTER site to reach a broader audience with ecological information. IMs envisioned a RDBMS capable of providing an organized long-term storage area for data and allowing the public and researchers to access data, metadata, and other information through a dynamic link to the SGS-LTER web site (1996 SGS-LTER proposal). These technological advances made it necessary for IMs to not only understand ecological sciences and current research projects, but also to be proficient in information technology, computer network and database administration, programming, and web site development. By building a team of IMs, we were able to garner expertise in several areas and work together with researchers to set new goals and meet new challenges (Figure 4).

This diagram illustrates the important functions in which the IM Team sees an opportunity to be involved. Components central to the IM system include non-spatial and spatial data and metadata management. Other pieces to the puzzle are part of greater efforts within the University, LTER Network, regional school districts, and the legislative branches of the government. As a team, we are able to cooperate with other ecological research and education and outreach organizations, agencies, and institutions, to manage, archive, and distribute data and information. The SGS-LTER project and LTER Network also benefits from retaining IMs that stay stimulated and valued in their jobs. We are able to attract part-time workers that want the opportunity to focus on their areas of interest and expertise and contribute to the work of the team to the benefit of the overall project.

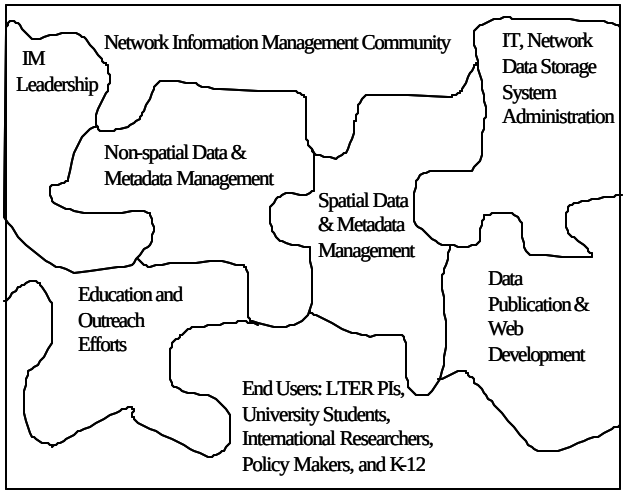

Figure 4. Information Management Team: current and proposed personnel components at SGS-LTER

Through the development of our team approach to IM, we have learned to balance the project's needs for cutting edge IT knowledge and development with efficient, more routine daily IM tasks, while maintaining a strong connection between the IM Team and the researchers. The community of IMs across the Network have to juggle a set of "healthy tensions" on a regular basis. These include balancing a top-down versus bottom-up approach; a site-centric versus network-centric focus; adapting software and other tools that are more appropriate for commercial, production-driven inquiries versus scientific, research-driven inquiries; needing standardization, but working in a network composed of a diversity of approaches; and overcoming the old-fashioned penchant for holding data in a centralized enterprise system versus a more modern enlightened approach of sharing data and metadata in a decentralized enterprise system. We are committed to sharing our experiences and history and becoming a mentor to new LTER sites as they plan their IM activities and organization.

Future Directions for Information Management at SGS-LTER and CPER

The SGS-LTER information management Team has identified areas in which we can become a leader in IM including managing models and their metadata, providing training and mentoring for other sites, and involving IM in our education and outreach programs. Managing metadata for models, controlling model versions, and documenting the range of validity, implicit assumptions, ecological context, and related long-term studies and data sets are challenging goals to meet. The IM Team would like to develop efficient ways to store, document, access and publish this type of ecological modeling information. We will collaborate with other modeling-intensive LTER sites and share results with the LTER Network and greater scientific community.

We also included an outreach and education coordinator on our IM Team. This is a timely opportunity to incorporate IM into Schoolyard LTER efforts and to impart a sense of the importance of IM to the next generation of scientists. We are able to work with existing education centers with ties to SGS-LTER, such as the University of Northern Colorado, Math and Science Teaching Center and the Center for Learning and Teaching in the West. IM will be a part of distance learning programs, professional development programs for teachers, and we will use the web as a tool for data delivery.

Our success as an LTER site depends on building partnerships with other LTER sites, as well as agencies, organizations, and institutions outside the LTER Network. We are currently enlisting the participation of ARS scientists and IMs in SGS-LTER research and information management activities. Researchers and IMs at SGS-LTER, CSU, and ARS are building tools to facilitate the collection and archival of project-level information and metadata in a RDBMS (Figure 5). This tool has proven to be useful and flexible enough to serve particular needs of managers at the SGS-LTER site, as well as the broader scientific community. People across the Network, academicians, agency personnel and policy makers can access this information for administrative, research, teaching, or synthesis research purposes with the use of active server pages (ASP) on our web site (Figure 1). Together with ARS staff, we were able to document project metadata in an accessible database that dates back to 1976. With this information, ARS site managers have the power to keep track of scientific experiments, treatments, land

use changes, and cattle stocking rates across the CPER site and to manage the land on which we conduct our research.

The IM Team also is excited about our active involvement with developing a "content standard" for metadata within the LTER community. This new tool, called Ecological Metadata Language (EML), will simplify data access (http://caplter.asu.edu/data/metadata/workshop012002.htm). The hierarchical design of our new data storage system and the new RDBMS containing metadata and project level information is similar to the design of the EML schema. We hope to be able to transfer metadata from our system to EML and take advantage of a remote server (e.g. Metacat) to reach a broader ecological community.

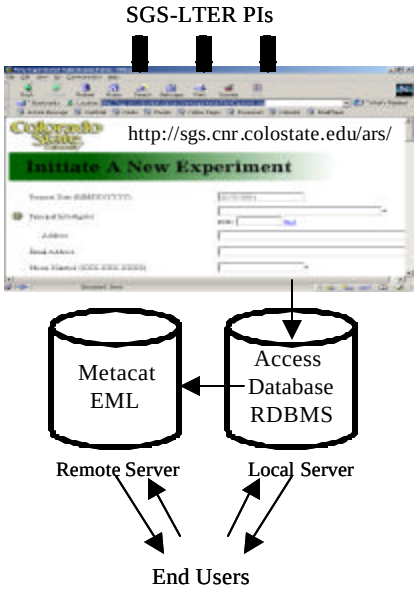

Figure 5. Web tool for researchers to contribute metadata directly to the SGS-LTER RDBMS.

Conclusions

The "Decade of Synthesis and Standardization" of metadata continues to be a pressing issue at both the site and Network level (Stafford, personal communication and Baker et al. 2000). The mission for LTER in 2002 has been reaffirmed. The central, organizing intellectual aim of the LTER program is to understand long-term patterns and processes of ecological systems at multiple spatial scales. To accomplish our mission we must gain a better understanding of the ecosystems which we study, by conducting cross-site, comparative research and synthesizing long-term data sets.

It is essential that LTER sites maintain a leadership role in the dissemination of information and continue to create well-designed, documented databases that are accessible on the World Wide Web. Our legacy of long-term data sets, as well as our understanding of IT implementation in information management is valued as an LTER asset. The Network has the expertise and the information to serve as a gateway for educators, policy makers, and greater scientific community to reach data and information to gain knowledge about a diverse set of ecosystems around the world. The site and Network's commitment to information management means that IM professionals will continue to participate in training and development exercises, and to share our experiences with a broader scientific community through outreach activities, such as this symposium (SCI2002 www.iiis.org/sci2002).

Acknowledgements

This material is based upon work supported by the National Science Foundation under Cooperative Agreement #DEB-9632852.

References

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

Burke, I.C. and W.K. Lauenroth. (1992) Enhancement of data management capabilities: a supplement to "BSR-9011659 LTER".

Brunt, J.W. (2000) Data management principles, implementation, and administration. Pp. 25-47. *In* W.K. Michener and J.W. Brunt. Ecological data design, management and processing. Blackwell Science Ltd., Oxford, UK.

Callahan, J.T. (1984) Long-term ecological research. BioScience. 34:363-367.

Chaffee, S. (1984) LTER Data Management reports: CPER Site report.

Kirchner, T.B. (1989) TIME-ZERO: the integrated modeling environment. Ecol. Modelling. 47:33-52.

Henshaw, D.L., M. Stubbs, B.J. Benson, K.S. Baker, D. Blodgett, and J.H. Porter. (1998) Climate database project: A strategy for improving information access across research sites. Pp. 123-127. *In* Michener, W., J.H. Porter, S.G. Stafford. Data and information management in the ecological sciences. A Resource Guide. Albuquerque, N.M.: LTER Network Office, University of New Mexico.

Stafford, S.G, P.B. Alaback, K.L. Waddell, and R.L. Slagle. (1986a) Data management procedures in ecological research. Pp. 93-113. *In* W.K. Michener, editor. Research data management in the ecological sciences. (Belle W. Baruch Library in Marine Science Number 16). University of South Carolina Press.

Stafford, S.G., M.W. Klopsch, K.L. Waddell, R.L. Slagle, and P. B. Alaback. (1986b) Optimizing the computational environment for educational research. Pp. 73-91. *In* W.K. Michener, editor. Research data management in the ecological sciences. (Belle W. Baruch Library in Marine Science Number 16). University of South Carolina Press.

Van Dyne, G.M. (1970) Progress, problems, and potential in the US IBP Grassland Biome Study. Memo October 14, 1970

Van Dyne, G.M. (1971) Grassland Biome, US International Biological Program. December Newsletter.