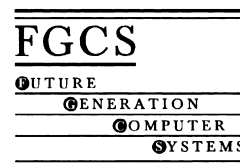




ELSEVIER

Future Generation Computer Systems 16 (1999) 21–28



A method for interoperable digital libraries and data repositories

John J. Helly*, T. Todd Elvins, Don Sutton, David Martinez

San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Drive, Bldg. 109, La Jolla, CA 92093-0505, USA

Accepted 11 February 1999

Abstract

Aside from the basic importance of metadata in documenting, the characteristics of data for reuse is the fundamental role it plays in the functioning of digital libraries and data repositories. Metadata provides both the content of the search catalogs and provides part of the basis for performing quality control on the source data. In this paper, we describe a method for a scalable and decentralized system of interoperable digital libraries and data repositories. The description includes transportable metadata format, a persistent naming convention for arbitrary digital objects and a protocol for the asynchronous distribution of metadata. We also include a description of an operational data repository based on these methods. ©1999 Elsevier Science B.V. All rights reserved.

1. Introduction

This paper reports a method for sharing metadata between digital libraries and data repositories. It is based on the notion of decentralized control with arbitrary redundancy between nodes in a manner analogous to 'mirroring'. This work has been performed in the development of a data repository for the publication of ecological data but with the specific intention of obtaining both generality and scalability in the design. Our approach is based on three key developments which we will discuss in turn below. The first is a transportable metadata format which we call BKM. The second is a persistent naming convention for arbitrary digital objects (ADO) and the third is a protocol for the asynchronous distribution of metadata.

2. Methods

Our objective has been to develop techniques for the controlled publication of digital scientific data. To do this we are developing systems which combine the functions of conventional libraries with digital library and data repository technology. The system reported here is designed for the publication of ADOs which, for the present technology environment, are defined as uniquely identifiable objects within a computer's file system. These may contain measurements, images, sounds or any other digitally recorded data including software and documents.

Our methods are based on the results from our earlier work on the environmental data repository for San Diego Bay [1] and track the emerging digital library technology. Our design for the San Diego Bay project was guided by the principal that our system should be maximally portable, inexpensive and simple with the intention that it be replicable at the level of the individual researcher. The resulting system is

* Corresponding author.
E-mail addresses: hellyj@ucsd.edu (J.J. Helly), todd@acm.org (T. Elvins)

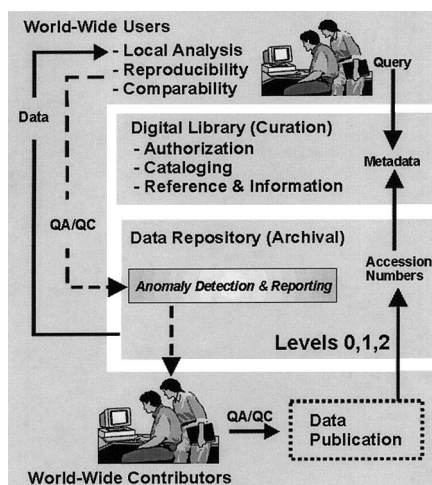


Fig. 1. Data flow interfaces supporting data publication.

workstation-based and is predicated on the Unix file system for its archive. For the project reported here, which we refer to as 'Caveat Emptor Ecological Data (CEED)' (<http://ecodata.sdsc.edu>), we are utilizing high-performance storage systems and servers located at the University of California, San Diego's San Diego Supercomputer Center (SDSC). The basic functional organization of the data repository and its relationship to digital library functions, data submitters and users are shown in Fig. 1. This figure illustrates functional interfaces in the operation of a data repository and also illustrates the concept of data levels and quality assurance/quality control (QA/QC) processing which will be described in a separate paper. While still replicable, this system is more complex but provides more extensive services. Our long-term goal is to achieve a family of client/server components which can interoperate to provide a seamless ability to develop and maintain individual and community collections of scientific data.

2.1. Digital libraries and data repositories

This is a period of intense activity in the evolution of libraries; involving librarians and computer scientists in new collaborations. Not surprisingly, there is considerable ambiguity in the conception of what future libraries will be and how they will operate. In developing the overall design of the CEED we elected

to differentiate a digital library from a data repository to clearly separate the role of curation, requiring domain-specific expertise, from archival, requiring computer resources and system administration expertise. We did this because it seems that digital libraries tend to emphasize metadata content while data repositories tend to emphasize data content. A recent description of the function of digital libraries emphasizes the distinction between the function of the library and the objects contained within the library:

'... The primary purpose of digital libraries is to enable the searching of electronic collections distributed across networks, rather than merely creating electronic repositories from digitized physical materials [2].'

A data repository stores, maintains, and enables access to digital objects, handles hardware, software, protocols, interfaces, content synchronization and related system-level infrastructure. This is a complex and expensive task primarily as the result of the rate of technology change and the general unreliability of computing and communication systems. In addition, data repositories must provide formal and public data management policy, support standards for interoperability and support rights management and the protection of intellectual property. The searching capability provided by digital libraries and the data management services provided by data repositories are both contingent on the existence of system and application information. For the CEED implementation, application metadata is that provided by the data submitter while system metadata pertains to ADO naming, access control and so forth. This information is all stored in metadata.

2.2. Transportable metadata

Metadata 'standards' are emerging throughout the scientific community as well as in the larger population of information managers and users. Some of the most active and authoritative work relating to metadata comes from the library community and is presently exemplified by the Dublin Core [3] and the Resource Description Framework (RDF) [4] efforts. Other efforts relating to spatial data [5] also have a long history. It is illuminating to reflect on the growth of this area before considering our approach.

The Federal Geographic Data Committee (FGDC) (<http://www.fgdc.gov>) was one of the first organizations to advocate a metadata standard specifically focused on geospatial data. This type of data is dominated by remote sensing imagery and map data such as digital elevation models (DEMs). Although widely criticized as unwieldy and rigid, their proposal was a milestone in providing a controlled list of metadata parameters for community consideration. The FGDC effort was driven by the proliferation of geographic information systems (GIS) technology within federal government agencies and the resultant need to facilitate data sharing between agencies. The standard also provided guidance to GIS software developers. Criticism of the FGDC standard stems from the large number of parameters it defines¹ and the fact that extensions to it were not allowed. Consequently, it was perceived by many as an overly complicated and closed standard imposed by government mandate on federal agencies and their contractors. Fortunately, the policy on extensions has been relaxed [6] and the concept of 'minimum metadata' (a 56-parameter subset) has emerged. These constructive changes are leading to wider acceptance but the large number of parameters continue to pose a daunting data entry task for individual scientific researchers. In any case the FGDC standard is widely considered definitive for geospatial data and many metadata developers appropriately adopt some subset of the FGDC parameters.

As part of an effort to preserve long-term ecological data [7] a group of ecologists and computer scientists (including one of us) collaborated in the development of a metadata standard for non-geospatial data [8]. This standard is designed for documenting data resulting from ecological research and captures metadata missing from the FGDC standard yet important to the ecological community. A key aspect of this standard is a three-tiered hierarchy for the acquisition of metadata. The hierarchy explicitly provides for metadata acquisition based on the data provider's purpose in providing the data for use by others. That is to, (1) provide data for use by an expert colleague; (2) enable spatial, temporal and keyword (thematic) searches to find the data; or (3) support auditing or peer-review of

the data². This hierarchy enables the data provider to determine how much or little effort to spend on metadata as a function of the intended reuse of the data. We have found this approach to be more readily accepted by individual researchers for whom the notion of data publishing is a new, unfamiliar and often burdensome task [9]. The same parameters were also organized into five broad classes (I–V) according to the type of information they contained (i.e., data set, research origin, data set status and availability, data structure, supplemental). The result was a two-way partition of intended use versus type-of-information.

The first implementation of the ecological metadata standard described above was done as part of the development of the web-site for publishing environmental monitoring data from the San Diego Bay (<http://sdbay.sdsc.edu>). Although more comfortable for individual scientists, this standard has considerable ambiguities which became clear during the implementation and preliminary use. In particular, some of the parameters initially considered to be part of the minimum set of metadata (i.e., providing data for use by expert colleague) were found to be unnecessary as part of the minimum set or more logically part of another class. We also found that it was useful to distinguish between the parameter name and the label used to display the parameter within a URL. However, despite great temptation to make unilateral changes to the organization of the metadata during the implementation, we maintained strict one-to-one correspondence of the parameters in the paper and the parameters in the HTML implementation as a means of configuration management. We found this to be of great help in establishing a review process with members of the ecological community which we describe below.

As a result of our experience with this initial implementation we came to a second metadata implementation which we dubbed BKM metadata format. We now refer to the San Diego Bay version as BKM V1.0 and the current version as BKM V2.x. The definition of this format can be obtained from <http://ceed.sdsc.edu>. From the start of the San Diego Bay implementation, we decided that a copy of the metadata should be packaged with its corresponding data in a tar file as

¹ 60 pages; 3–9 parameters per page.

² The full complement of parameters through the third level of the hierarchy is 63.

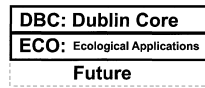


Fig. 2. RDF-compatible BKM structure with Dublin Core (DBC), Non-geospatial metadata for ecology (ECO) packages and room the growth. Acronyms indicate the package in the RDF sense.

part of an ADO. Therefore, an ADO always contains at least two files: the data file and the metadata file. For the metadata file to be maximally useful we also decided that it should be formatted as a 'flat, ASCII file' with a published format and semantics to enable it to be easily and unambiguously interpreted by a human using a conventional ASCII editor. These conventions have the obvious additional benefit of being well-suited to the construction of translators designed to convert from BKM format to another target format. As a result, the BKM format serves as a convenient mechanism for the transport of metadata between digital libraries and data repositories in addition to helping the data user make effective use of the data.

BKM V2.x format is in third-normal form [10] to ensure the unique identification of every record in the metadata file and the records are structured to be consistent with the RDF scheme for formalizing, generalizing and sharing metadata. It is a transportable, flat, ASCII, metadata file format with records structured into [name, value, label] triples. The triples are organized into modules or packages identified by a three-letter acronym coded into the name attribute of the triple. This is done to the modular RDF scheme and we encode RDF packages using the three-letter code (e.g., DBC, ECO) at the beginning of each record. This is shown conceptually as concatenated or stacked sections within a BKM metadata file in Fig. 2 and exemplified in Table 1. For example, 'dbc' corresponds to Dublin Core parameters, 'eco' corresponds to ecological parameters. This structure was designed to facilitate the sharing and transport of metadata by simplifying machine-parsing while conforming to emerging standards and preserving the ability to read the file contents directly in a conventional ASCII editor.

BKM parameter names accommodate a two-level hierarchical structure with nodes and subnodes corresponding to major metadata categories such as those described in [8]. The parameter names are composed of fields as well as sorting keys (i.e., pkgserial, classe-

rial) to organize the parameters within packages and classes, respectively. The fields within the parameter name are delimited by the underscore character ('-') and the comma as shown in Table 1 while the name, value and label attributes of the schema are delimited by a semicolon. The semicolon was used to simplify parsing under Java. Underscores separate fields of text contained in the fully-qualified parameter name and the comma separates fields used only in composing the presentation view.

The fields coded within the parameter name are as follows:

pkg	dbc for Dublin Core, eco for ecological.
pkgserial	sequence number within a package.
class	A, B, C, D corresponding to the extent of metadata. A is considered minimal.
classserial	sequence number within the class.
node	hierarchical node name.
noderpt	repeat group number within the node. For example, file 0, file 1.
subnode	subordinate hierarchical node name. Node value is repeated if no subordinate node exists.
subnoderpt	repeat group number within a subnode. For example, variable 0, variable 1.
parameter	metadata parameter name.
label	presentation label.
default	default parameter value.

Each parameter name is unique and its value is allowed to be of arbitrary length. To accommodate this feature the rows of the BKM files are also sequentially ordered with sequence numbers contained in the parameter name to enable a parameter to be composed of multiple BKM records. This is needed to support repeating parameters such as those relating to a given file when multiple data files are being combined into a single ADO. This is also done to accommodate arbitrarily long parameter values such as narrative text using the limited length records of current operating systems.

The BKM format is currently undergoing a review as part of an Ecological Metadata Working Group sponsored by the National Center for Ecological Analysis (NCEAS; <http://www.nceas.ucsb.edu>) and Synthesis and the San Diego Supercomputer Center (SDSC; <http://www.sdsc.edu>). Other participants in the group include the Long-term Ecological Research

Table 1
BKM V2.2 file example with a parameter name is bold>

```

Bkm_metadata_version,'V2.2','Metadata Version'
# BKM Metadata Master Version 2.2 1998-06-30
# Format of entries in this file are:
pkg_pkgserial_class_classserial_node_noderpt_subnode_subnoderpt_parameter,'defaultvalue','label'
# dbc indicates Dublin Core parameters
# eco indicates Modifications from Michener et al., Ecological Applications, 2/97
# Some original eco parameters have been superseded by dbc parameters
#
# Send comments to hellyj@ucsd.edu
#
dbc.10000_A_5000_dataset.0_dataset.0_resourceidentifier;'TBD';'Resource Identifier'
dbc.1000_A_1000_dataset.0_dataset.0_title;'TBD';'Dataset Title'
dbc.2000_A_2000_dataset.0_dataset.0_creator;'TBD';'Resource Author or Creator'
eco.1000_A_3000_dataset.0_dataset.0_address1;'TBD';'Author/Creator: Address (1)'
eco.2000_A_4000_dataset.0_dataset.0_address2;'TBD';'Author/Creator: Address (2)'
eco.3000_A_5000_dataset.0_dataset.0_address3;'TBD';'Author/Creator: Address (3)'
eco.4000_A_6000_dataset.0_dataset.0_phone;'TBD';'Author/Creator: Phone Number'
eco.5000_A_7000_dataset.0_dataset.0_email;'TBD';'Author/Creator: Email Address'
dbc.3000_A_8000_dataset.0_dataset.0_subject;'TBD';'Subject Keywords'
dbc.8000_A_8000_dataset.0_dataset.0_resourcecetype;'TBD';'Resource Type'
dbc.6000_A_9000_dataset.0_dataset.0_contributor;'TBD';'Other Contributor'

```

Network Office (LTER; <http://www.lternet.edu> as well as representatives from the federal government and ecological research stations. NCEAS is developing an XML³ version of the same set of metadata parameters and it is our intention to evolve toward both implementations through a joint configuration management approach. This will result in both the as yet hoped-for benefits of the XML representation as well as the benefits and portability of the 'flat ASCII' representation. Likely future modifications will involve the separation of FGDC parameters into a package within BKM and the evaluation of Z39.50 package⁴.

2.3. Uniquely identifying persistent digital objects

One of the most troublesome problems in developing and maintaining a data repository is the establish-

ment of a system of unique and persistent names for the digital objects. There have been many schemes proposed as solutions. One group has proposed the notion of handles [11] and there have been others with similarly novel names. Significant efforts currently underway focus on the creation of Universal Resource Names (URNs) but current definitions do not specify a particular implementation although there are useful recommendations resulting from these efforts (<http://www.w3.org>). Consequently, we are left to determine the structure of our particular implementation of a URN but have attempted to accommodate the recommendations in our approach.

We have established a system we think of and refer to as accession numbers drawing on the card-catalogue analogy. These accession numbers have a structure that is illustrated in Fig. 3. It is useful to note that this accession number structure is easily extensible to a federation of data repositories as illustrated in Fig. 4 One interpretation of the Repository of Record and Userid fields could be that of publisher and author. The main features of this encoding is that an ordinary file system sort will result in the ordering of all files based on the subfields (repository, userid, original datetime, version datetime) from left-to-right. Consequently, all versions of the same ADO will be conveniently listed

³ Extensible Markup Language (XML) is a language for publishing Standard Generalized Markup Language (SGML) on the world-wide-web. The figure of XML is not entirely certain so these two approach (XML and 'flat ASCII'), while complementary, are intentionally being independently maintained.

⁴ As stated <http://lcweb.loc.gov/z3950/agency/>, the Z39.50 standard, 'Information Retrieval (Z39.50): Application Service Definition and Protocol Specification,' is represented as both ANSI/NISO Z39.50 and ISO 23950.

Repository of Record	Userid	Original Datetime	Version Datetime	File Type
CEED	u0010	980211.230010	980317.000020	tar

Persistent Name

Fig. 3. *Accession Number Structure*. In this illustration the CEED data repository name is coded as part of the persistent name. Note that this convention can be augmented on the left to participate in external associations of data. The date format is being modified to accommodate the YYYYMMDD format.

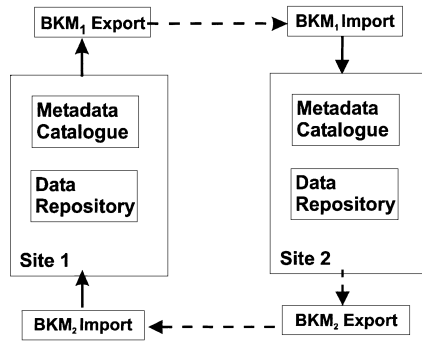


Fig. 4. Metadata distribution using BKM.

together and all ADOs belonging to a given user will likewise be grouped together.

2.4. Asynchronous distribution of metadata

The problem of ensuring the completeness of a specific search for data is rife throughout the world-wide-web at present. This is due to the lack of a complete description of the search space largely resulting from the volatility of URLs. The same problem is inherent in the publication of data through digital library technology. Since the search-space is defined by the universe of metadata, the ability to form the union of relevant metadata within a domain is essential to the completeness of a search. Clearly there must be reliable method of sharing (i.e., publishing) metadata in a systematic and reliable way. BKM format provides a general and simple means of doing this as depicted in Fig. 4.

Fig. 4 illustrates the use of a BKM formatted file as the mechanism for transport across data repository and digital library sites enabling the synchronization of metadata catalogues across a federation of interoperable sites. Each site can, for example, publish

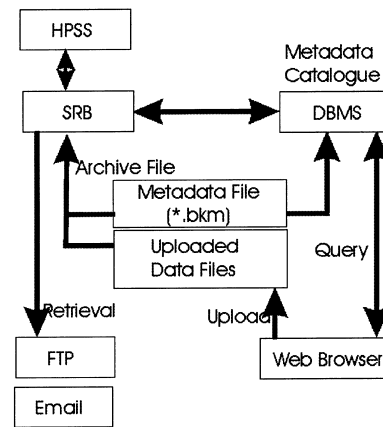


Fig. 5. CEED system architecture depicting major functional components and interfaces and the use of BKM files.

its entire metadata catalogue or parts of it for other sites to retrieve on demand using conventional methods such as ftp or http. As a result of the structure of the BKM format, each site can independently interpret the BKM parameters into its particular implementation of a metadata catalogue. While we do not encourage arbitrary and independent interpretation of the parameters, this flexibility acknowledges the existence of important differences in the way in which different disciplines use and construe the same information contained in metadata.

3. Results

Using these concepts work has been completed on a system for the controlled publication of 'Caveat Empor' ecological data [12]. These are data which have not been peer-reviewed but which are offered for use by the data submitters to facilitate the research of others and the progress of collaborative ecological science. This site is the initial implementation of the methods described in this paper. The architecture for the system is shown in Fig. 5. The overall features of this architecture include a web-based user interface providing for user registration, search and retrieval data contribution and contributor-based access control. A WWW browser is used for the user-interface. A DBMS is used to maintain the metadata catalogue which is built from contributor-generated metadata first written to a BKM file which is packaged with the

contributed data into the ADO named with the convention described above. The BKM file is then used as input to the DBMS used to maintain the metadata catalogue. The storage resource broker (SRB) provides a scalable, operating-system level interface to arbitrarily decentralized storage systems including the HPSS system for the initial system [13–15]. Conventional internet services of email and HTTP and used in the notification and download process for retrieval of data. Access control is provided through the DBMS services as well as the SRB. File-level control is provided to enable individual data submitters to control access to any given digital object they have published in the repository.

One of the key features of this architecture is the separation of the interactive functions (i.e., queries, data selection, data submission) from the retrieval and delivery of data to the requesting user. There are unpredictable latencies in most, if not all, archival systems which must accommodate delays as large as those required to manually load a new tape into a robotic storage system. There are also the unpredictable failures of network components to contend with. By separating the delivery of data from the querying and submission of data we provide the user with greater reliability as well as avoiding the inconvenience of tying up a web-session waiting for inherent system latencies. Once a retrieval is completed, the user is notified by email that the data he or she requested is available for downloading via ftp (or http) through a selectable URL embedded in the email message itself. We also maintain a log of users who have retrieved any particular ADO so that they can be notified if a new version of the ADO is published and to inform the ADO provider of users who have copies of the ADO. Data providers find this feature to be reassuring especially in relation to concerns about intellectual property rights and attribution for the use of their data.

4. Summary and conclusions

We have described three key components to the construction of a set of interoperable digital libraries and data repositories: a transportable metadata format, a persistent naming convention, and a simple method of sharing metadata between. These results are derived from approximately three years of work involving a

variety of experiments. From this we have distilled a few key points which we list here. These should be considered with an awareness that our implementation has been developed to specifically address the needs of the individual scientist wanting to cooperate in the sharing of research data. From this perspective we clearly see the continuing need for a:

- user-selectable set of metadata parameters which can be progressively completed based on the data providers available time and motivation,
- persistent naming convention which is unique within the data repository and clearly associates any given digital object (ADO) with its provider as well as related versions of the object,
- simple, implementation-independent implementation of structured metadata (e.g., BKM) which can be directly read and interpreted by an individual scientist,
- mechanism for bundling metadata with the ADO in an archive file to ensure it is delivered with the object when copied, and
- mechanism for non-interactive data retrieval and delivery to a user to accommodate the latencies inherent in the archival storage system.

We encourage others who may develop such systems to consider these points in their development efforts. For these methods to be useful it is essential that a critical mass of user interest be crystallized in a manner similar to that which produced the current system of libraries and more recently the Internet and the WWW. This will clearly take some considerable time and experience with other approaches and the further development of publishing tools such as XML as well as policies regarding intellectual property which have not even been discussed here. Clearly, this work represents only one approach which could serve as a basis for conventions and perhaps standards. Other experiments should be done. While our approach may or may not be adopted by others, in a general way the issues addressed in this work are fundamental to the development of a robust means of controlled publication of digital scientific data.

References

- [1] J. Helly, San Diego Bay Project, 1998, <http://sdbay.sdsc.edu>.
- [2] B.R. Schatz, Information retrieval in digital libraries: bringing search to the net, Science, 1980.

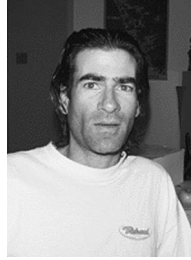
- [3] PURL, Dublin Core Metadata, 1998, http://purl.org/metadata/bublin_core.
- [4] <http://www.w3.org/RDF>, 'Resource Description Framework (RDF) Model and Syntax Specification, W3C Working Draft 19 August 1998', W3C.
- [5] <http://www.fgdc.gov/metadata/constan.html>, 'Content Standard for Digital Geospatial Metadata (CSDGM)', Federal Geographic Data Committee, 1988.
- [6] Federal Geographic Data Committee, M.A.H.W.G., Content Standard for Digital Geospatial Metadata, Reston, Virginia, US Geological Survey, 1998.
- [7] K. Gross, E. Allen, C. Bledsoe, R. Colwell, P. Dayton, M. Dethler, J. Helly, R. Holt, W. Michener, N. Morin, S. Pickett, S. Stafford, A. O'Neill, C. Pake, Report of the Committee on the Future of Long-term Ecological Data (FLED), 1995.
- [8] W.K. Michener, J.W. Brunt, J. Helly, T. Kirchener, S. Stafford, Nongeospatial Metadata for the Ecological Sciences, *Ecological Applications* 7(1) (1997) 330–342.
- [9] J. Helly, New concepts of publication, *Nature* 393 (1998) 107.
- [10] J.D. Ullman, Principles of Database Systems, Potomac, Maryland, Computer Science Press, Rockville, MD, 1980.
- [11] <http://www.handle.net>, The Handle System, 1998.
- [12] J. Helly, CEED: Caveat Emptor Ecological Data Repository, 1998, <http://ecodata.sdsc.edu>.
- [13] C. Baru, R. Frost, R. Marciano, R. Moore, A. Rajasekar, M. Wan, Metadata to support information-based computing environments, IEEE International Conference on MetaData 97, September 1997.
- [14] C. Baru, M. Wan, A. Rajasekar, W. Schroeder, R. Marciano, R. Moore, R. Frost, Storage Resource Broker, <http://www.npaci.edu/DICE/>.
- [15] A.R.C. Baru, 1998, A Hierarchical Access Control Scheme for Digital Libraries, Third ACM Conference on Digital Libraries (June).



Dr. Helly is responsible for the Earth Systems Science program at the San Diego Supercomputer Center. He has a wide range of experience in ecology, statistical analysis and systems architecture for large-scale computing and communication systems. Helly has a PhD in Computer Science, an MS in Biostatistics from UCLA and an MA and BA in Biology from Occidental College. hellyj@ucsd.edu.



T. Todd Elvins is a Principal Scientist at San Diego Supercomputer Center. He has just finished his Computer Engineering PhD at the University of California, San Diego and his research interests include perceptually-based user interface design, data and information visualization, web-based imaging, and computer graphics. todd@acm.org, <http://www.sdsc.edu/todd/>



Don Sutton is a research scientist at the San Diego Supercomputer Center. He works with the Earth System Science, ESS, group on projects which bring advanced computational technologies being developed at the center to basic researchers in ESS. For the most part this includes integrating database technology, ESS computational models, visualization tools, and advanced web architecture. He has a PhD in Bioengineering from UCSD and has worked at Scripps Institution of Oceanography and the Naval Research and Development Lab (SPAWAR) in San Diego.



Mr. Martinez holds an MS from the University of California, San Diego and has worked for three years at the San Diego Supercomputer Center as an application programmer. Mr. Martinez is an expert at web applications and cgi programming.