Building the National Virtual Collaboratory
for Earthquake Engineering Research

NEESgrid

Technical Report NEESgrid-2004-43

www.neesgrid.org

(Report draft: 1.0)

Last modified: June 30, 2004

Summary Report on NEESgrid's Data Curation Summit

Kincho H. Law
Professor of Civil and Environmental Engineering
Stanford University
Stanford, CA 94305
Email: law@stanford.edu

Feedback on this document should be directed to law@stanford.edu.

Acknowledgment: This work was supported by the George E. Brown, Jr. Network for Earthquake Engineering Simulation (NEES) Program of the National Science Foundation under Award Numbers CMS-0117853.

1 Introduction

To develop an agenda on NEESgrid Data Curation and related issues, a Summit meeting was held in Chicago on March 18, 2004. Coordinated and sponsored by the NEESgrid System Integration team, the summit brought together experts in library information science, earthquake engineering, data infrastructure, and data curation, to forge a forward-looking plan needed to improve the NEESgrid data usage and curation. This report briefly summarizes the discussions in the meeting and outlines the data curation needs for NEESgrid.

NEESgrid is intended as a distributed virtual "collaboratory" for earthquake experimentation and simulation. This collaboratory will allow researchers to gain remote, shared access to experimental equipment and data. The system infrastructure is designed to support data repositories, data sharing, and access. Tools will be released that enhance the sharing, access, and utilization of the NEESgrid data repository, and thus enhance efficient communication among the researchers in earthquake engineering.

From a broader perspective, NEESgrid has the potential to provide a tighter linkage between research and practice, and between the earthquake engineering community and the public, through outreach and education. To do so, the NEESgrid effort will need to be extended to encourage the community (1) to utilize and access the repository, (2) to learn about the experiments and the results, and (3) to discover new information from the repository populated with experimental data. Data curation, the process of compiling, organizing, and cataloging project information, as well as the information about the data, plays a very important role in facilitating the usage and dissemination of the archived experimental data and information.

The NEESgrid team is very much interested to see how far the current system can be used to support data curation, to learn the current state of practice in the archiving of scientific and experimental data, to define the requirements for the curation of NEESgrid data, and to investigate the possibilities for extending NEESgrid to serve the experimental researchers, earthquake engineers, scientists, and the public. The objective of the data curation summit meeting was to initiate a dialogue among the experts in data repository developments, library information science, and earthquake engineering and to help define NEESgrid's needs for data curation. The appendix includes the meeting agenda and a listing of the participants. The meeting included presentations on NEES and current efforts on NEESgrid, presentations by the participants, and a brainstorm discussion session. The presentations and written discussion materials by the participants can be found in the NEESgrid website http://www.neesgrid.org/curation/.

2 Background and Issues

Interest in the fields of library science and information archival, as well as in data access and curation issues [1] has increased dramatically in recent years. Figure 1 shows a functional model of an Open Archival Information System, which may serve as a reference model for discussion purposes [2,3]. In simple terms, an OAIS serves to facilitate efficient dissemination of digital data and content archived in a repository. The goal of NEES's repository is similar. As for NEES, the producers are the experimenters and researchers who produce the data to be ingested into an archival storage system (repository). The data management system supports typical access functions such as searching, viewing, integrity control, and retrieving the data. The access functions serve to receive requests, check privileges, and generate and deliver responses to the "customers"; the customers, in this case, are the researchers, practitioners, educators, students, product manufacturers, and, potentially, the general public.

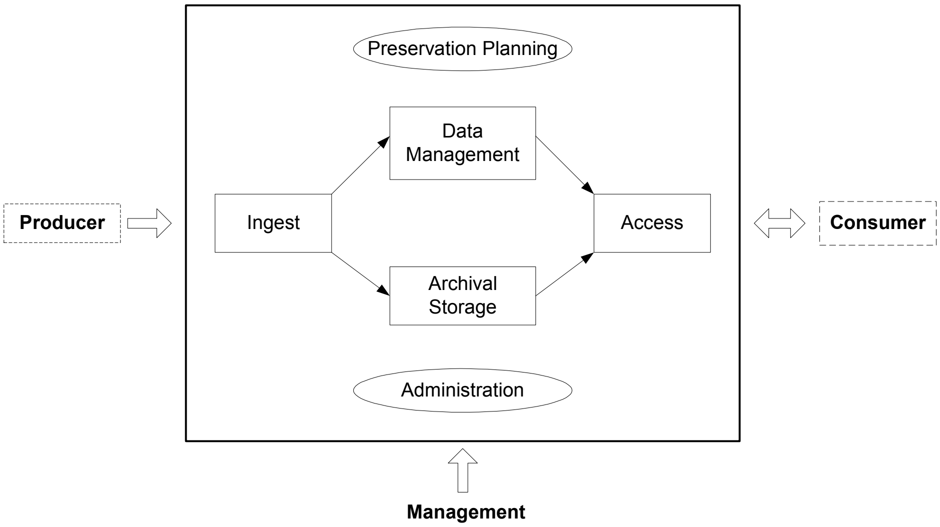

Figure 1: OAIS Functional Model [2,3]

Current development of the NEESgrid system focuses primarily on data modeling, data ingestion tools, the data repository and tools that *directly support experimental activities*. For example, as shown in Figure 2, the current data modeling efforts deal primarily with data and metadata specifically to support experiments. To date, data curation, management, and preservation have not been a main focus of NEESgrid activities.

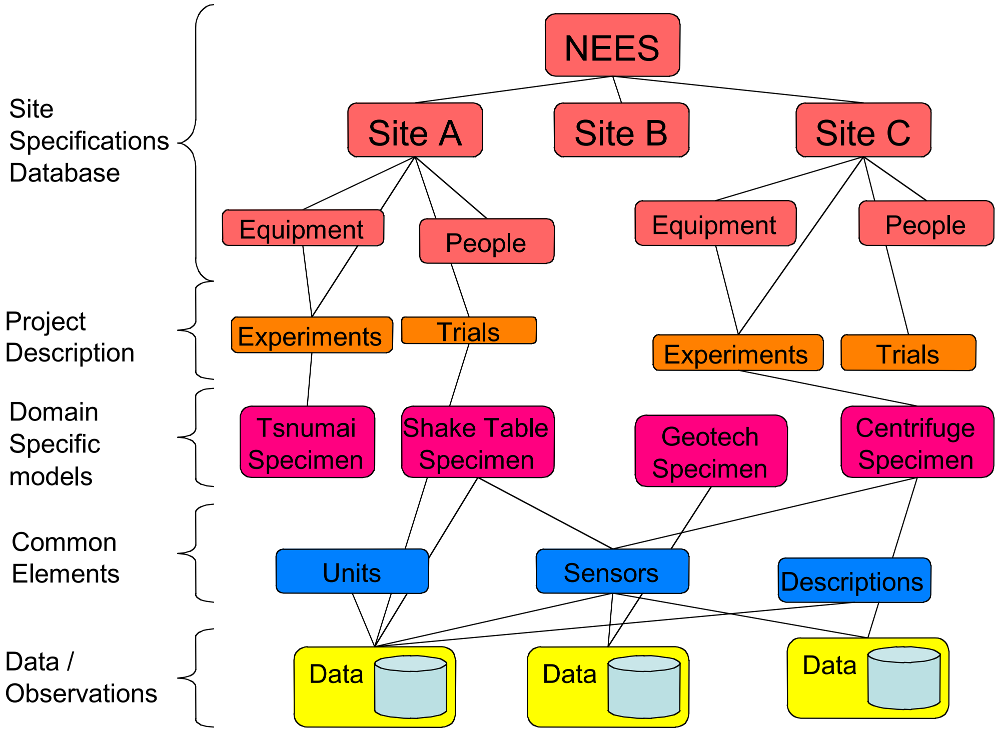

Figure 2 – Overall Data Model for the NEESgrid in Supporting Experimental Activities

The following include a list of questions and issues posed in the NEESgrid Data Curation Summit meeting intended to discuss the policy and the technology needs to facilitate the data production/generation, data management and preservation, and the consumption and use of the data.

Questions related to Objectives, Administration, and Policies:

1. *What are the desirable objectives for the data repository and usages for the NEESgrid?*
2. *What audience should the NEESgrid data repository target?*
3. *What does data "curation" mean in the earthquake engineering context?*
4. *What is the scope for the data curation and access that NEESgrid should deal with?*
5. *What are the long term and short term objectives for the NEESgrid data repository?*
6. *What are the administrative (ownerships) and preservation (initial versus ongoing) policies that need to be considered?*
7. *What types of data should be preserved for use by the "consumers"?*
8. *To what extent should NEESgrid be involved in developing tools for data curation and data analysis and exploration?*
9. *Who will pay for the initial and the ongoing cost of curating data and of managing archived, curated data?*
10. *How do other scientific communities approach this problem? What are some good examples of curated resources which reflect the work output of a community as diverse as the Earthquake Engineering community?*

11. *How should the NEESgrid repository handle a possibly diverse set of data models? Or should we standardize on a model and force all of the data in the repository to conform to the "one" model? If we allow flexibility in models, what are the best practices in tracking how models evolve over time?*

Questions related to Technology and Tools:

12. *Are there existing tools that are being used in production that could be deployed for the NEES central repository? Who can we talk to to gain additional information?*
13. *What are the common workflows that are used in other efforts of this nature and how those workflows might be applied/modified for use in NEESgrid? What tools are commonly used for this workflow?*
14. *What need to be done in order to extend current system to support data management and access for the "consumers" using NEESgrid?*
15. *What are the available tools that can help defining metadata to support data curation and make the data easily accessible, understandable, usable?*
16. *Are there any available data curation and access tools that could handle the data types for NEESgrid?*
17. *What level of data curation can be achievable in the short term (6 months - 1 year)?*
18. *What level of data curation should be targeted in the long term (5+ years)?*

3 Data Curation for NEES

According to the UK Digital Curation Centre (see http://www.dcc.ac.uk/what.html), data curation involves "The actions needed to maintain digital research data and other digital materials over their *entire life-cycle* and over time for current and future generations of users. Implicit in this definition are the processes of digital archiving and preservation but it also includes *all the processes needed for good data creation and management, and the capacity to add value to data to generate new sources of information and knowledge*." That is, curation implies well-planned active management of information and involves the production, conservation, preservation and access of the data. The management of data must ensure that the people to whom the data is relevant can find the data. Furthermore, curation needs to ensure supports of data/information reuse and facilitate generation of new information and knowledge from the data. The following summarize some of the issues and recommendations brought forth in the meeting.

3.1 Data Generation/Ingestion (The Production Phase)

NEESgrid tasks have dealt primarily with the "data production" phase. Current developments and tools, including project browsers, an e-notebook, data models, will facilitate experimenters and researchers generating and ingesting data into the data repository which will in turn allow users to browse and possibly search the data/information about a project or a specific experiment. These developments aim to support the basic NEES's policy that "no *data* generated by NEES should be lost."

The selection of data to be ingested and preserved is a difficult issue. Specific to NEES, not only the generated/sensed data are important, but also the process to generate all the data is of great importance. Earthquake experiments also include observations such as sensor output, video, and metadata, all of which need to be preserved. Data "appraisal", i.e. the selection of data to be preserved and in what format, must be part of the data ingestion phase and the policy of the curation process.

Another important issue is that the data ingestion process needs to be automated as much as possible. The process needs to have the ability to capture the contextual information and the structure of data. A consistent and logical naming convention/structure (or namespace), directory structure, and metadata to help automate ingest, search, order and persistent naming must be defined. A structure map, which includes the logical and physical structure of the objects, needs to be defined and standardized. A standard data verification process is needed to guarantee the integrity of data. Creation date and a digital signature may be needed to verify the data sets (original versus derived, etc.). Quality control, including the status of the data ingested (unedited, edited), is important to ensure the data in the repository can be trusted.

One important note is that data ingestion must link closely with data policy. Policy issues related to data ingestion may include "when the data is considered ready for ingestion?" and "what kind of data to be saved -- raw data, curated data, or processed/calibrated data?". It may also be preferable to have policy to define two or more levels of archival -- one level of acceptance for just "storing" quality data, and a much higher, much more qualified second level as a goal for completely indexed, database-style searchable data/metadata. In other words, what level of data should be saved in order to reproduce higher level information and knowledge?

3.2 Data management and preservation

Data management involves a persistent organization to coordinate the experimental centers as well as a centralized center that has the capability to manage the data – with the business function of data management and archival. Dedicated communication between data management staff and those producing the data is important. Replication and backup strategies (ready to mitigate risks) need to be defined.

The data saved and stored in the repository should always be available to the people who use the data. The repository design and deployment need to be scalable and sustainable in the long term. As data may be changed or amended with new information over time, revision control and versioning are important issues. One policy issue has to do with the authority over the data, i.e., who has the right to make changes to the data. Furthermore, consistency of data needs to be maintained, in that a persistent namespace structure is fundamentally important for data ingestion, management and preservation. It is necessary to document as details as possible and store the data associated ("encapsulated") with the experiments even if the data format may not be well understood (such as different video streaming, access and display formats) so that all data will be captured.

Data management should also provide support of an audit trail over the lifecycle of each data object. The audit trail information should be tied closely with data model and data policy, which needs to provide information about what model, what policy, and where in the lifecycle the data was created. Audit trail and quality control are also important for building confidence on the data.

One special issue in data preservation is proprietary data format. This issue arises when data is generated and stored in particular data format using specific commercial software. As much as possible, it is desirable that the data be captured in some non-proprietary format and encapsulated as one file or byte stream that is part of the digital object. In addition, it may be necessary to save the software itself with the data. A good mechanism is needed for supporting data transformation (for example from RDBS to file systems and vice versa) and data migration (to be manipulated by a newer version of software). A further complication is found in the heterogeneity of NEES data and documentation, which often include a variety of data types from different platforms (and equipment) including drawings, text, sensor readings, relational databases, etc. Different types of data may require different curation strategies.

Even with excellent file standards, policies, audit trails, etc., files will evolve and new types will arrive over time. Translation software (from one format to another) will have to be maintained. Development and maintenance of such "ancillary" software will require significant efforts even though the immediate benefit to data generators (experimenters and researchers) of such software may not be clear.

At the present time, the wide variety of data (shakers, waves, stress tests, etc.) and experimental methods require the initial NEES archive effort to focus on a limited data management scope. However, tools should be developed and made available for researchers/experimenters to improve quality data/metadata gathering, which will lead to more usable archived data. Policies, guidelines and roadmap for data sharing and archiving need to be integrated and implemented with a robust data management and preservation strategy to support access.

3.3 Data access and consumption

As noted by a workshop participant, “Almost by definition, research data is a ‘one each’ type of problem, and computer interfaces for such minimally-reproducible problems are non-trivial.” Curation cannot be divorced from use. Providing easy access to meaningful and useful NEES data is an important goal.

There are fundamental needs to provide access mechanisms even when the experimental data is to be accessed years after an experiment. One purpose of the NEES data repository is to support long-term usage and potentially the discovery of new knowledge. One question is how to use the saved data to support knowledge discovery. For NEES researchers, they should be able to query and access past experiments that related to the one that he/she is contemplating. Some of the example questions may include:

- Who did the experiment? How can this person be found now?
- What business rules were in place at the time, what data was captured and where it is stored?
- What was the status of the facility at the time?
- What was the sensor status -- type, calibration, maintenance?
- What other experiments, trials, computations, etc., have used the data?
- Is the experiment referenced in an online journal?
- What NEESpop and NEESgrid version, DAQ application version, etc., were used to originally capture and ingest (import) the data?
- What application software and versions were used for analysis?
- For each individual object, is this the originally encoded (e.g., via QuickTime) or is it translated and encoded? Can the uncompressed original dataset be viewed?
- Where does each data object fall in time and space within the experiment? Is it possible to virtually reconstruct the experiment?

Easy access to the information must be supported. However, secure access to data repository must also be provided; access control is also needed. Another issue is accessing based on roles, i.e. presenting data to different audiences: K-12 students, general public, and researchers. It will be very desirable if NEES data can be integrated with complementary data sources from other communities such as IRIS [9]. Such integration not only encourages cross-disciplinary research, but also can be part of a collaboration model for long-term sustainability of the NEES repository.

3.4 Data policy

As discussed, many of the issues for data ingestion, data management and data access, are intimately related to the NEES's policy. The current policy by NEES's Data Sharing and Archival Committee (DSAC) is that "Data must be ingested in the repository so it will not be lost. Data must be retrievable from the repository over time so it will not be lost." The current efforts by the DSAC have been focusing on policies (1) to establish rules and guidelines to insure data archiving and sharing; (2) to preserve intellectual property; (3) to implement data quality assurance procedures; and (4) to implement electronic publications for data preservation and release. In addition, there are other policy issues that worth further consideration.

- Data archiving should include guidelines on what, how, and when data is to be archived. Related to this point is a question about when the data actually becomes an archival object and can therefore be "deposited" in the repository. What data is to be stored locally and what is stored centrally? What is the time lag between local storage and central storage? Documentation is important if there is any divergence between the two. For example, is there a clear path that can trace back to the local data required for anything archived locally?
- There is a need of a data sharing plan to include minimum requirements. Business rules need to be defined and enforced so that all data produced should be able to be analyzed and handled according to the business rules. The rules should be clearly identified whether they are mandatory, recommended, or optional.
- Authorization and access privilege need to be defined and authenticated. Who has the authorization to modify the data, to reuse data, to remove data, etc.?
- There is a strong need to evaluate/identify requirements for near and long term users. Possibly the initial heaviest users of archival data will be the researchers, investigators, and the students since they should find the easily accessed, well organized, and documented storage of their own data useful. Policies are also needed for long term data archival and access so that new knowledge and research may be generated from the saved data. Plans are also needed for an evolving archive interface (improved ease-of-use, K-12 outreach, etc.).
- The data policies and guidelines cannot be viewed as a static entity, but must be regularly reviewed (e.g., on a semi-annual basis by the DSAC) to ensure that the objectives of NEES are met and that the community is well served.

4 Summary and Recommended Action Plan

The design and development of the data curation process, data management strategy and access tools is a very important goal since a repository is only as good as what can be done with the data for science and the education. Digital data curation is an active area of research and development. The participants have recommended further investigations and possible collaborations with other related efforts; these include the DSPACE project [5], research center on digital curation [6], the digital preservation program by the Library of Congress [7], the Skyserver project [8], the IRIS project [9], the Fedora program [10], ICPSR on social science research archive [11], FGDC on federal government mapping program [12], and others.

Data and metadata management and curation are very important part of the NEESgrid. However, as noted by one participant, "The efforts required to archive research data are often underestimated…. Custom interfaces are expensive to design and the free exchange of information between researchers in tight funding times is not always as simple as one would assume…" Long term commitments and "business" model are needed to sustain and facilitate optimal usage of the data in the repository. Many issues have been discussed in the meeting. Some of the recommended immediate actions that may worth undertaken are summarized as follows:

- While "the wide variety of data (shakers, waves, stress tests....) and experimental methods will likely require the initial NEES archiving effort to focus on a limited data management scope," tools are needed "to help researchers/experimenters to improve quality data/metadata gathering, which will lead to more usable archived data." There is a continuing need to further develop (data and metadata) models and tools to facilitate data importing, sharing, access and utilization of the NEESgrid data repository. Data should be tagged with usage-related metadata describing its validity, quality, provenance and status. Access control and policy need to be defined and implemented as part of the data management scheme.
- "Possibly the initial heaviest users of archival data will be the original investigators and students since they will hopefully find the easily accessed, well organized and documented storage of their own data useful." Current development focuses on the data generation and ingestion phase of the curation process. Data services that will illustrate potential use of the data repository need to be developed in order to encourage community use.
- Data policy issues affect all phases of the data management and curation process. "There is a strong need to evaluate and identify the needs for near and long term users." The works by the Data Sharing and Archiving Committee are commendable and their activities are of paramount importance towards defining the roadmap for data management and preservation strategy. It is particularly important that the policies and guidelines be coordinated and integrated with the tool development efforts. Specifically, there is a need for a (short and long term) implementation plan of policies and guidelines in the data management and curation development.

In summary, the NEESgrid effort has been successful and has achieved its initial goal in providing tools to support collaborations and experimentations. From the broad data curation perspective, however, as pointed out by a meeting participant that the "system (is) built backward: sensor data collection, NEES, data management, ..." There is a need for the NEES and the earthquake engineering community at large to continue working and defining a road map to help access and use of the data in the repository (both short- and long-term), leveraging available tools, developing new tools, as well as establishing management and administrative schemes and policies. Relevant technologies should be assessed and evaluated to establish a technical strategy. From the technical strategic perspective, the NEESgrid community needs to address whether data curation should be considered as a diverse set of activities or a tightly control process. Could a single, standardized tool be developed to handle the heterogeneous features of NEES's facilities and tests? What investments should be made to extend NEESgrid's existing tool set to support the curation features? The Data Curation Summit Meeting has provided an initial overview of the issues and recommendations that need to be further investigated by the NEES community.

5 Acknowledgments

This summary report would not have been possible without the many insights and discussions by those who participated in the NEESgrid Data Curation Summit meeting. The participants gave generously of their time and expertise, and have provided many slides and written documents (see website http://www.neesgrid.org/curation/) on which this report has drawn and replicated. Although this report has attempted to summarize the discussions at the Summit meeting, it does not necessarily represent the views of the the NEESgrid SI team, the meeting participants or their organizations, nor should it be construed to represent any consensus statement or shared set of findings or recommendations. Last but not least, the author would like to extend his thanks to Joe Futrelle for his valuable comments on the report and to Cristina Beldica, Patty Kobel, Bill Spencer, Chuck Severance and Grace Agnew for their help in preparing for the meeting.

References

1. www.openarchives.org (OAI – Open Archives Initiative)

2. G. Agnew and A. Alcts, Developing a Metadata Strategy, 2/2003. (available at http://gondolin.rutgers.edu/MIC/text/how/metadata_agnew.pdf)

3. Reference Model for an Open Archival Information System (OAIS), Consultative Committee for Space Data Systems, CCSDS 650.0B-1, Blue Book, 1/2002. (available at http://ssdoo.gsfc.nasa.gov/nost/isoas/)

4. J. Gillilan-Swetland, Setting the Stage, in Introduction to Metadata – Pathways to Digital Information,. 2000. (available at http://www.getty.edu/research/institute/standards/ intrometadata)

5. Bass, et.al. DSpace—A Sustainable Solution for Institutional Digital Asset Services – Spanning the Information Asset Value Chain: Ingest, Manage, Preserve, Disseminate. Internal Reference Specification- Functionality. Version 2002-03-01. (available at http://libraries.mit.edu/dspace-mit/technology/functionality.pdf)

6. Digital Data Curation Research and Development Center (proposal), Johns Hopkins University, 2004 (available at http://dkc.mse.jhu.edu/RD-Data-Center.pdf).

7. National digital information infrastructure and preservation program, a Collaborative Initiative of the Library of Congress, 2000 (available at http://www.digitalpreservation.gov/)

8. Gray, J., Szalay, A.S., Thakar, A.R., Stoughton, C., vandenBerg, J. , "Online Scientific Data Curation, Publication, and Archiving," Technical Report, MSR-TR-2002-74, Microsoft Research, Microsoft Corp. July 2002.

9. The Incorporated Research Institutions for Seismology (see http://www.iris.edu)

10. The Fedora™ Project, "An Open-Source Digital Repository Management System"(see http://www.fedora.info/)

11. Inter-University Consortium for Political and Social Research (ICPSR), (see http://www.icpsr.umich.edu/.)

12. Federal Geographic Data Committee (see http://www.fgdc.gov/).

Appendix

NEESGrid
NEES System Integrator Data Curation Summit
Chicago O'Hare Hilton
March 17-18, 2004

AGENDA

Wednesday, March 17

6:30 – 9:00	**Dinner**

Thursday, March 18

8:00 - 8:30	**Continental Breakfast**
8:30 - 8:45	Welcome, Overview, Introductions -- Bill Spencer
8:45 - 9:30	NEESgrid Data Technology Overview -- Charles Severance
	Describe the overall architecture, goals and objectives of the NEESgrid data activity.
9:30 - 10:15	NEESGrid Data Curation from the SI Perspective -- Kincho Law
10:15 - 10:30	**Break**
10:30 - 11:15	NEESgrid Data Curation from the NEES Consortium Perspective – Andrei Reinhorn
11:15 - 12:00	Data Specialist Perspective on NEESgrid s Data Curation Needs (15 minutes each)
12:00 - 1:00	**Lunch**
1:00 - 2:30	Data Specialist Perspective on NEESgrid s Data Curation Needs (15 minutes each)
2:30 - 3:30	Discussion and drafting of review and recommendation document
3:30 - 3:45	**Break**
3:45 - 5:00	Continued discussion and drafting of review and recommendation document
5:00	**Adjourn**

NEESGrid
NEES System Integrator Data Curation Summit
LIST OF PARTICIPANTS

Participant	Organization	Email Address
Grace Agnew	Rutgers, the State University of New Jersey	gagnew@rci.rutgers.edu
Murtha Baca	Getty Information Institute	mbaca@getty.edu
Bruce Barkstrom	NASA Langley Research Center	b.r.barkstrom@larc.nasa.gov
Cristina Beldica	NCSA	cbeldica@ncsa.uiuc.edu
Rick Benson	IRIS Data Management Center	rick@iris.washington.edu
Fran Boler	UNAVCO, Inc.	boler@unavco.org
Jim Eng	University of Michigan	jimeng@umich.edu
Ronald Jantz	Rutgers, the State University of New Jersey	rjantz@rci.rutgers.edu
Joanne Kaczmarek	University of Illinois	jkaczmar@ux1.cso.uiuc.edu
Patty Kobel	NCSA	pkobel@ncsa.uiuc.edu
Kincho Law	Stanford University	law@stanford.edu
Myron McCallum	UNAVCO, Inc.	myron@unavco.org
Viswanath Nandigam	SDSC	viswanat@sdsc.edu
Gokhan Pekcan	University of Nevada-Reno	pekcan@unr.edu
James Peng	Stanford University	junpeng@stanford.edu
Andrei Reinhorn	University of Buffalo	reinhorn@buffalo.edu
Beth Sandore	University of Illinois	sandore@uiuc.edu
Charles Severance	University of Michigan	csev@umich.edu
Wayne Shiver	UNAVCO, Inc.	shiver@unavco.org
Bill Spencer	University of Illinois	bfs@uiuc.edu