TR-2005-[ID]

An Assessment of Data Curation Issues for NEES

Kincho H. Law, Jun Peng and Peter Demian

Stanford University

Last Modified: 2005-09-30

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 Number CMS-0117853.

Table of Contents

List of Figures

Acknowledgement

This report is written with the support by the George E. Brown Jr. Network for Earthquake Engineering Simulation (NEES) Program of the National Science Foundation under Award Number CMS-0117853. The authors would like to thank Dr. Joy Pauschke, the Program Manager of NEES at NSF, and Prof. Bill Spencer of University of Illinois at Urbana-Champaign, the Principal Investigator of NEESgrid System Integration (SI) Project, for their encouragement and support. The authors would also like to thank Dr. Anke Kamrath, the Principal Investigator of the NEES CyberInfrastructure Center (NEESit), for arranging the meeting with the specialists in digital library and data archiving at San Diego Supercomputing Center on February 24, 2005. The authors have greatly benefited from the two Data Curation Summit Meetings on March 18, 2004 in Chicago, Illinois and on July 28-29, 2005 at University of California, San Diego, sponsored respectively by the NEESgrid SI Team and the NEESit Center. Data curation is a rapidly developing field. With the nature and the diversity of opinions on the subject, the materials and discussions expressed in this report are solely based on the views of the authors and do not necessarily represent the views of the participants attended the two Data Curation Summit Meetings or their organizations, the NEESgrid SI Team, the NEESit Center, NEESInc or the National Science Foundation nor should this report be construed to represent any consensus statement or shared set of findings or recommendations at the two Data Curation Summit Meetings or by any organizations.

Executive Summary

The NEES (George E. Brown's Network for Earthquake Engineering Simulations) infrastructure is intended as a distributed virtual "collaboratory" for earthquake engineering experimentation and simulation. When fully developed and deployed, this collaboratory will allow researchers to gain remote, shared access to experimental equipment and data. The system infrastructure is designed and tools are being developed to support the archival and access of research project data. If properly facilitated, the archived information can potentially be shared by a broad audience, thus providing linkage and communication among the researchers, between research and practice, and between the earthquake engineering community and the public. Data curation, a process to compile, organize, and catalog the project data and the information about the data, will play an important role in facilitating the access and use of the archived experimental data and project information.

Over the years, there has been a significant amount of research and development in the digital library and data preservation areas. Many of such efforts could be relevant to the NEES data repository and curation effort. Data curation spans the entire data lifecycle from data production to consumption and is an activity not only for managing the data but also for promoting the use of data, ensuring quality for data reuse and supporting research and knowledge discovery. So far, the effort with the NEES infrastructure has dealt primarily with the basic process of storing and archiving data in a repository. An interim data strategy has been designed to achieve its initial goal to upload and store the project data. Other efforts towards the development of data models for supporting researchers' needs have also been pursued. To ensure that the archived data can be accessed by the researchers and, potentially, the public, a well-designed curation process will need to be put in place.

The importance of the data curation effort to ensure longevity, sharability and accessibility is common to many disciplines, from social science to oceanography. Interests in digital data preservation and curation have grown in recent years. Many research programs and centers for data curation have been launched in the US and Europe. The purpose of this report is to assess the issues, needs and requirements relevant to the NEES program. Related literature on data curation is reviewed to extract useful concepts that may be applicable to the NEES data repository and curation. The report is structured following the functional model by the Open Archive Initiative (OAI) (see http://www.openarchives.org), which includes the basic activities involved in the data curation process -- data ingestion, data management and preservation, data archiving and administration, and data access. The report reviews the current state of research and practice in these areas and highlights the issues deemed most relevant to the NEES program. Decisions about data curation involve not only technologies and technical developments but also economic, legal, social and organizational considerations. Collaboration among the researchers, IT developers and management team is necessary. It is recommended that a clear vision, administrative plan and policy, roadmap, commitments for short (< 5 years), medium (5-10 years) and long (> 10 years) terms be defined by the NEES management and the earthquake engineering community. This report is written in the hope that it will further stimulate discussions on the works ahead to develop a viable, cost-effective curation plan for the NEES research program.

1 Introduction

The NEES (George E Brown's Network for Earthquake Engineering Simulation) infrastructure is intended as a distributed virtual "collaboratory" for earthquake engineering experimentation and simulation. The collaboratory allows researchers to gain remote, shared access to experimental equipment. It is also expected, through the NEES research program, significant amount of valuable data and knowledge will be generated. The system infrastructure is designed and tools are being developed to support data archive and access.

The NEES research program has the potential not only to support research but also to provide linkage between research and practice, and between the earthquake engineering community and the public. The data effort with the NEES infrastructure has so far been focused on the basic process of capturing the experimental and simulation data in digital form and storing them in a repository. An interim data strategy has been designed to achieve its initial goal in establishing a data repository for storing the project data. Other efforts towards the development of data models for supporting researchers' needs have also been pursued. In order to establish a repository supporting data sharing and access by researchers and the public, there is a need for the NEES management, IT development team and the earthquake engineering community at large, to define a road map for both the short-term and long-term archival strategies and the use of NEES data.

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*." (See http://www.dcc.ac.uk/what.html.) That is, curation implies well-planned active management of information and involves the production, archival, preservation and access of the data. The management of data must ensure that the people who are interested in the data can find the data. Data curation, a process to compile, organize, and catalog the project information and the information about the data, will play a very important role in facilitating access of the archived experimental data and project information. Furthermore, curation needs to ensure supports of data/information reuse and facilitate generation of new information and knowledge from the data.

Scientific research activities, such as the NEES research program, generate and accumulate large quantities of information that need to be archived, and maintained in a trustworthy environment and kept accessible for long period of time. Such information are of value to scientists and public alike – they are worth retention for future use to support research, practice, education, public policy and planning. Government agencies are increasingly concerned about the preservation of such valuable information and preventing the potential loss of research investments. As noted by Larry Brandt, NSF Program Coordinator of DIGARCH, "Digtial preservation is of central importance for scientific data … As a society, however, we are creating more and more information which is digital in its original form (NSF Press Release 05-074)." Over the years, there have been many significant developments in the digital library and data preservation areas. However, there are very few reliable methods to systematically manage digital content over the data life cycle. Further complicated the problem is that digital content is typically fragile, volatile and highly dependent upon hardware and software. Any information in digital form is vulnerable for long term risk. Obvious problems, such as obsolescence of software and hardware and versions and format changes, easily make archived digital data inaccessible. Digital data,

even stored in the simplest form as bit streams, are in danger for retrieval. Digital data archiving and preservation, which takes full consideration of access and retrieval, have been of significant concerns for quite some time (Lesk 1992; Rothenberg 1995; Chen 2001b; Buneman and Foster 2002; Ray et.al. 2002; Berry et al. 2003; Gray et.al. 2005). In addition to the technical problems, there are administrative, workflow, legal, economic, organizational and policy issues surrounding the problem of digital data archiving and publishing. Unlike traditional paper-based storage, the management of digital data must take into consideration lifecycle process and the means by which the data is generated, captured, transmitted, stored, maintained and access. Both technical and non-technical issues cannot be separated in developing an economically viable scheme for data curation.

The purpose of this report is to review some of the current works related to data curation and to stimulate the discussion on the needs and requirements for NEES's data curation effort. Related literature on data curation is reviewed to extract useful concepts that can be applied to the NEES research program. The report is structured according to the functional model for Open Archival Information System (OAIS 2002), which includes the activities involving data generation/ingestion, data management, archiving and preservation, and data access and control.

This report is organized as follows: First, an overview of data curation and the OAIS functional model is introduced. The discussions then focus on the various issues related to data generation and ingestion, data management and preservation, and data access and collaboration. Some important administrative and policy issues worth consideration are also briefly reviewed. This report concludes with a brief summary of the report and highlights the items that may worth further review and discussion.

2 Data Curation – An Overview

Scientific activities such as observations, experiments and computer simulations gather and/or generate scientific data that are saved and published. Traditionally, the producers of the data are also the primary keepers of the data. There is no general techniques for keeping the data as long-term archives or for efficient retrieval from those archives (Buneman et al. 2002). Scientific data, collectively, represent the intellectual capital of a community. The collection contain not only the digital entities (the data) that comprise the digital holdings of the community, but also the context (the metadata characterizing the data) required to interpret and manipulate the digital data and collection. During the curation process, the content is analyzed to identify features within the data. The features are labeled and stored in the form of descriptive metadata as part of the context of the data set. Scientific data collections thus serve as the repository for the information that a scientific discipline has assembled (Chen 2001a).

To facilitate use of scientific data, data curation is a critically important process. Scientists assemble and publish data to share with other scientists. Scientists also devise new research projects based upon prior collected research results and derive new scientific findings. Educators use the material as sources for preparing teaching materials. Government agencies and public organizations use the data to develop policies. The use of the word "curation" in the fields of digital archiving and records management is still nascent. Therefore, the meaning and the scope of data curation are still open for discussion. The term "curation" builds on the understanding of the word "curator", which is somebody who keeps something for the public good (Lord and Macdonald 2003); for example, a historical museum curator is responsible for selecting artifacts to be preserved and displayed in a museum for the sake of history.

Thus, data curation involves activities for selecting, managing, preserving, and adding value to the digital collection of data.

For a collection of digital materials, much of the data curation activities are keenly related to digital library and data publication. A digital library generally has a domain focus and its collection often serves a specific purpose (for example, art, science, or literature). Also, it is usually created to serve a community of users. A typical digital library holds a collection of digital objects, which can be electronic books, journals, documents (e.g., pdf files, HTML pages), and multimedia materials (such as pictures or images, tapes or video files, etc.) which are stored in some locatable repositories. Besides the digital data objects, a digital library also holds a collection of metadata structures, such as catalogs, guides, dictionaries, thesauri, indices, summaries, annotations, glossaries, etc. A scientific data publication system will need to support ingestion of the digital objects and the metadata about the objects, querying of metadata catalogs to identify objects of interests, and potentially the integration of information across multiple data collections. From a user's perspective, a digital library system can be viewed transparently as a collection of widely distributed, autonomously maintained data repositories. Services are provided to support activities such as document summarization, indexing, collaborative annotation, format conversion, bibliography maintenance, and copyright clearance. A library uses quality control in the sense that all its material is verified and consistent with the profile of the library. The material is filtered before it is included in the library, and also its metadata is usually enriched with annotation and categorization. The digital library also has the responsibility to ensure protection of information of enduring value for access by present and future generations. Preservation includes regular allocation of resources for persistence, preventive measures to arrest deterioration of materials, and remedial measures to restore the usability of selected materials. As communication technology advances, a digital library could become interoperable with other digital libraries, forming a web of ubiquitous libraries accessible by the users (Moore et al. 2005).

Perhaps the simplest framework to discuss the issues of data curation is the functional model of an Open Archival Information System, OAIS, as shown in Figure 1 (OAIS 2002). 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 retrieval of 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. Note that the issues of data ingestion, management, archival and access are interrelated in the overall data curation framework. It should be kept "in mind that the OAIS is intended as a reference model rather than a system design model…..the functions or processes….do not necessarily correspond directly to the functional modules of a system that would implement that model (Rothenberg 2000)." Nevertheless, the OAIS model provides a unified framework to examine some of the fundamental issues that may be of relevance to the NEES data repository and curation effort.

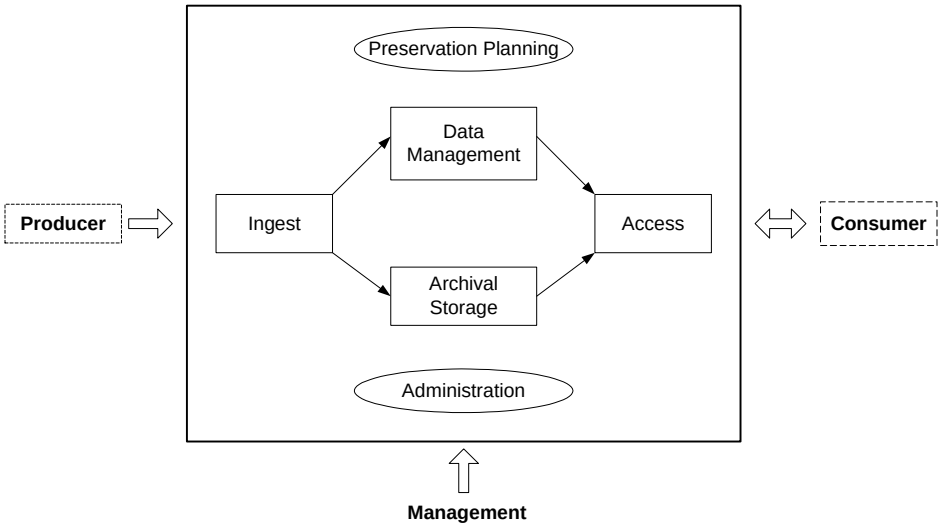

Figure 1: OAIS Functional Model (http://www.openarchives.org/)

As noted in the OAIS model, data curation is characterized as the whole process of data generation and ingestion, data storage and management, and data access and consumption. The following sections discuss some of the key issues related to this process. Among the activities of data curation will also include administrative policies and quality control, and implementation features such as authentication, access control, audit trials, replica management and persistent identifiers (Rajasekar and Moore 2001).

3 Data Generation/Ingestion (The Production Phase)

The first phase of the data curation process concerns with the data generation and ingestion tasks. Experimenters and researchers generate, process and ingest the data into the data repository which will in turn allow users to browse and possibly search for data/information about a project or a specific experiment.

3.1 Data Selection

Scientific activities, such as the NEES research program, produce large volume of data. There are primary research data (raw data such as sensor readings), secondary (derived or processed data) and tertiary materials (such as research results and findings, often in the form of published papers and reports). Typically the volume of primary data is larger than secondary data, whose volume may also be larger than that of tertiary data. Traditionally, only tertiary data and part of secondary data are saved and, at times, maintained and made accessible to the other researchers. With the advance of storage and network technologies, it becomes feasible and economical to store and manage primary scientific data (Moore 2000). Scientific disciplines are starting to assemble primary source data for use by other researchers beyond those who gathered the data. Among the major reasons to keep and maintain primary, secondary and tertiary research data include: to facilitate reuse of data for new research, to retain unique experimental or observational data that is impossible or very expensive to re-create, to validate research results, and even to support the use of data for teaching and by the public (Gray et al. 2002). Even though storage systems could well have the capability to accommodate all the data, data

selection policy may still be needed to ensure the digital data collected are of lasting social, cultural, educational and research value (Hodge 2000).

The selection of data for ingestion and long term archival and preservation is not a trivial problem. The Data Sharing and Archival Committee of NEES has established a policy regarding data archival. The purpose of the policy is to ensure and facilitate full and open access to quality data archived in the NEES data repository. The basic NEES's policy is that "no *data* generated by NEES should be lost. … the data documentation must contain sufficient information about the data and its creation to ensure that any qualified user could reproduce experiment or simulation… (NEES-DSAC 2004)." In principle, all researchers should comply with this policy to the extent possible. Whether the policy is realistic or achievable for all experiments and projects (independent of sizes and resources) and by all NEES researchers may worth further deliberations. Specific to NEES, not only the generated/sensed data are acquired, but also the process to generate all the data may be needed to "reproduce the experiments and simulations". Earthquake experiments and simulations include video, application programs, documentations, metadata, observations such as sensor output, and many other data, all of which worth to be preserved. Even a simple experiment could generate enormous amount of data if all documentations, raw and synthesized data, and results are to be stored. Furthermore, digital documents (data file, video, etc..) and simulation results often require specific hardware, software or equipment (and their versions) to make them truly usable. Many of these tools are of proprietary nature and could be expensive to acquire. Even if standardized tools can be selected or acquired for specific tasks (which is unlikely in a research environment), as the technology and environments become obsolete, such digital materials may become unusable (Rothenberg 1995). What is considered appropriate and "sufficient" information about reproducing an experiment or simulation will likely be different among different researchers. The time imposed upon researchers to select and ingest "sufficient and reproducible" data could also be quite substantial. The data archival policy may clarify the types of data (including the process or workflow, software and hardware specification, etc.) to be preserved and archived and, more importantly, appropriate tools that help fulfill the policy and requirements are provided to assist the data selection and ingestion process.

3.2 Metadata (or meta-metadata)

Metadata (data about data) definitions and documentation are among the most critical tasks in the data curation process (Gillilen-Swetland 2000; Baca 2000; Agnew and Alcts 2003). They are essential to provide detailed information about a project, its motivation and the results and serve as the bridge for data ingestion and data access. To facilitate the data ingestion process, standards for defining metadata and directory structures need to be established. The metadata should be capable of capturing the contextual information and the structure of data in addition to the description of the data content. Metadata, if defined properly, can facilitate organization of the experimental project information, support various view definitions and help search and retrieval of relevant information from the project database. To decide the type of metadata information for different research projects, however, could be a difficult task. Deciding the granularity of the metadata – in other words, how much is enough and how much is too much – is non-trivial and may well be different among disciplines, fields of interests, research projects and the use of the data repository.

There have been numerous proposals and research in defining metadata structure in digital library (Baldonado 1997a, 1997b; Baca 2000; Hillmann2005). Metadata schemas and standards, particularly

for supporting education, exist. Examples include Dublin Core (DC) (ISO 2003), IEEE's Library Object Metadata (LOM) (2002), and others (e.g., see http://metamanagement.comm.nsdlib.org/IntroPage.html for links to other common metadata standards). Metadata allows precise description of data sources in small well-defined set of data elements without including the data sources themselves (Hillmann 2005). Metadata information designed for digital library and archive typically focus primarily on high level information such as Title, Creator, Subject, Description, Format, Media Types, etc., which are useful for search facilities quickly and accurately find the objects and contents of interests to the user. However, there is no "one-size-fit-all" metadata solution. The richness of the metadata structure and amount of elements depends on expressive power, cataloging efforts and users' needs for a specific discipline. To facilitate access in a specific domain (e.g., earthquake experiments) and encourage interoperability with other related fields (e.g., seismology), vocabularies, taxonomy/ontology and thesauri are needed to provide the necessary knowledge and organization structure for information retrieval and mining (Hovy 2003; Lau 2004; Lau et al. 2005a, 2005b; Salton and McGill 1983). The taxonomy and vocabulary could also be important for searching related experiments, simulations and projects on specific subject that are archived in the repository. Information retrieval tools can be developed and employed to assist in establishing some of the metadata by extracting terms and related information from project files (typically structured or semi-structured documents) stored in the repository. Nevertheless, experience has shown that building a viable metadata structure is an iterative process requiring participation of domain users (Heath 2005).

Metadata structure designed for a digital library and archive is not necessarily the same as a data model designed for a database or a data repository. A data model is a specific description of the type of information and relationships between the information needed for a specific problem domain. It provides a conceptual view or definition for the database or repository to meet the needs of the applications and users (Codd 1982; Ullman 1989; Date 2004). For the NEES data modeling effort, the reference data model developed for an earthquake experimental and simulation project on Shake Table and Large Scale Structures testing is intended to facilitate the design and implementation of a database or repository (Peng and Law 2004a, 2004b). Both an object structured model (expressed in Web Ontology Language, OWL (OWL 2004)) and a relational model (expressed in terms of an entity relationship model (Chen 1976, Teorey 1990)) have been put forth. Started with 5 basic data entities - Site Specific Information, Project and activities, Equipment Setup, Apparatus information and Data Elements – the data model contains about 20 object classes or relations that were deemed important by the experimental researchers to classify and represent the data sets. To illustrate the use of the reference data model, a facility was built to retrieve and categorize the archived data sets (in either an OWL file structure or a MySQL database system) according to a report structure typical of an experiment (Peng and Law 2005). The purpose is to demonstrate and validate specific and detailed information, for examples, on excitation, experimental setup and sensor measurements, etc., which are often of interests to earthquake engineering researchers, are represented in the reference model and retrievable from a repository implemented according to the model.

Cautiously, the data model is named a reference model since it is designed to provide an organization of the data according to specific types of earthquake experimental projects. The model may not have sufficient details for certain users and applications. Neither is the reference data model intended for specific implementation of the data repository. For instance, Fenves and McKenna (2004) proposed extensions to the reference data model for simulation and Swift et al (2004) incorporated extensions for centrifuge experiments. Other researchers are interested to include detailed descriptions of testing

specimens (Kim 2004). Industrial standards and product models (such as SensorML (Botts 2002) for GPS or GIS related applications useful for field tests, STEP (ISO 1994), CIMSteel (Garas and Hunter 1998), and ifc (IAI 2002) models for specimen descriptions) could be append to the reference data model if desired. Other NEES data models have also been proposed (Kutter et al. 2002; OSU 2003; Warnock 2005). In practice, a data model, irrespective to its complexity, should be transparent to the users while tools for data entries, views and applications are developed to assist users according to their needs. It is undoubtedly that depending on the needs of the users, equipment sites, the experiments' needs and the earthquake engineering community as well as the tools and implementation, the data models will (and should) evolve and change accordingly. Various scientific disciplines are also developing its own markup language and data model for describing its domain specific information (see for examples, AIML for astronomical instruments (Sall and Ames 1999), CML for chemistry (Murry-Rust et.al. 2001), EML for ecology (EML 2005) etc..

Although the metadata and the project data content stored in the repository can be mapped and referenced, the purpose of the metadata structure and the data model is different. The two could share and have overlapped information. Some of the metadata can be extracted from the project data content stored in the repository. It is not uncommon that a scientific digital library adopts standard metadata elements (such as Dublin Core, LOM, etc..) together with domain specific metadata and data model to describe scientific data sets. It is important to separate the metadata structure from the archived data repository. Libraries may store the metadata in their online public access catalogs. Publishers may store the metadata in a bibliographic or citation database. Electronic journals are tagged with headers for title, authors, affiliation. In a distributed archive environment, metadata may be stored centrally, with actual data objects distributed throughout the network (similar to most of the web archived system). Metadata, in some way, represents "a canonical form for a class of digital objects that, to some extent, captures the essential characteristics of that type of object in a highly determined fashion (Lynch 1999)." A data model, on the other hand, represents an "organizational" structure of the data objects to be stored in a repository and is designed for specific domain application, in this case, for earthquake engineering experiments and simulations.

3.3 Data Ingestion

A typical earthquake engineering experimental project (which includes the entire process of planning, simulation, experimental tests, synthesis of results, etc.) can easily produce hundreds of gigabytes (if not terabytes) of data and thousands of data files. Even a simple experimental event can generate hundreds of files. Robust technologies are needed to automate the data and metadata ingestion process (Jaja 2005). It is of utmost importance that data ingestion be automated as much as possible to relieve the burdens off from the researchers so that they can focus their time on the research tasks at hand. Tools should be developed and made available to help researchers to archive their project information and documents as the project progresses (as opposed to treating data ingestion as a post-project activity), even if it means to be for temporary storage. Such tools may include "drag-and-drop" utilities, e-notebook for experimental tasks, easy-to-use tools for data entries about sensors and measurements, direct ingestion of sensor measurement data and site equipment data (together with appropriate metadata) to a data repository.

Data and metadata models play an important role in facilitating the development of data ingestion tools. These tools must be able to automatically decompose a data collection into individual objects and then

ingest the digital objects into the repository. The tools also should have data processing capabilities, such as automatic association of meaningful metadata according to the metadata structure and definitions with original scientific data (obtained from the experiments, simulations or observations), extraction of digital objects from proprietary formats, automatic mining of attributes used to describe each data objects, and cross-platform to work for a wide class of data from word processing documents, drawings to video. A consistent, logical naming convention, persistent directory and database structure, and metadata can be helpful to automate ingestion and access of the data objects.

Besides the development of easy-to-use data ingestion tools, incentives are important to encourage scientists and researchers to prepare and ingest data into the repository. The primary interests of most researchers (in the NEES program or other areas) are in the activities of the project, from the design and execution of the experiments to synthesis of the results. Researchers then publish the results to publicize and gain recognition of their research. NSF and NEES have well defined best practice policies and requirements regarding data ingestion for their funded activities. The practice assumes the researchers (data producers) will voluntarily prepare the data suitable for archiving based on their self interests or promotion, professional norms, and "own conscience". Experience and reality from other fields however say otherwise. As noted by Hedstrom (2005), "Compliance with requirements (with formats, metadata, documentations, etc..) are rare." Preparing well-choreographed digital materials ready for archive is a time consuming and, often, laborious process. Researchers (data producers) should be encouraged and rewarded for their efforts to prepare archive-ready data sets that meet the requirements.

- What would be the appropriate incentives to encourage NEES researchers to willingly ingest the project information that will subsequently be used by others?
- What incentives are available to encourage compliance? How would compliance be enforced?
- How effective are the available tools to enable researchers to prepare for "archive-ready" data and to help them ingest data?
- Could current practice and strategies be designed to provide incentives to encourage researchers to ingest data voluntarily?

Providing positive incentives may well reduce archiving costs and produced high quality data stored in the repository for future use.

4 Data Management and Preservation

Software solutions and digital repository systems to assist the management and delivering of digital materials are now begun to emerge. Among the popular solutions are the Fedora (Lagoze, et.al. 2005; see also http://www.fedora.info/) and DSPACE (Bass et al. 2002; see also http://www.dspace.org) projects which are under active development. These projects have developed tools to capture, index, store, support access control and deliver digital materials to remote users over the web or as a web service. To endure long term values of a data repository, however, data management and data preservation remain the key issues of concerns. Data management involves a persistent organization and strategy to coordinate data ingestion and archival storage mechanisms and to support access and usage of the data. Data preservation is important to ensure protection of the information for access by current and future users (Conway 1990). Long-term digital preservation is a difficult but important challenge in data curation. Beyond the technological issues, data management and preservation must be

viewed from both the perspectives of the users and the custodian of the data. Policies will involve researchers, experimental centers as well as the data center that has the capability to manage the data to define the business functions for data management and use.

4.1 Repository Design and Storage

One consideration in the overall data management scheme is the issue of short term (temporary) versus long term (permanent and persistent) storage and federated versus centralized data repositories. Data federation leaves the source data (or some of the data) and control at the production (which may be the equipment or researchers') sites while provides support for a global view of the data sets. The federated environment imposes significant demands on the data repositories for management control and consistency. Depending on the sophistication of individual researcher, equipment sites and available resources, a federated environment may be difficult to enforce for quality and sustain. Current NEES repository focuses on a centralized permanent data repository paradigm, which is, relatively speaking, easier to manage, but scalability and performance issues may arise in the long run.

The Data Sharing and Archival Committee of NEES has defined an overall dataflow scheme and requirements for researchers to archive the data (NEES-DSAC 2004): "Data … must be stored as preliminary data in an initial repository as they are recorded. The investigators must submit structured data to the permanent repository within six months from the end of experiment or simulation, after proper quality evaluation, integration with documentations and other necessary preparation for curation. The data submitted to the permanent repository will be curated and archived in the permanent repository from which these data will be made accessible to the public." This specification implies that the initial repository would likely be resided at the researchers' home system or at the equipment site. In addition to the data ingestion tools to the initial repository, "bulk" upload of the data sets from the initial repository to the permanent storage will be needed. Consistent metadata structure, including persistent namespace structure, would greatly facilitate the upload process. An alternative is to establish temporary storage at the centralized site that would allow individual researchers to upload research data and documents. Whether the project data is to be temporarily stored at the researcher's, equipment, or centralized repository site, the issues that need to be dealt with involve the assignment of data upload responsibility, management of the process, versioning control and quality control in the permanent data archival process.

If the repository is to be maintained for long term use, backup strategy, where data is replicated and copies are maintained at separate locations, should not be overlooked (Cooper et al. 2000; Goldberg and Yianilos 1998; Kubiatowicz et al. 2000; Liskov et al. 1991). Repository design and deployment need to be scalable and sustainable for a long period of time. For long term archival, accidental deletions, natural disasters, and bankruptcy of the institution holding the collection could happen (Cooper and Garcia-Molina 2001). Data management thus involves selecting appropriate scheme to ensure reliability of data storage and has put in place strategies ready to mitigate risks.

Further consideration for repository design should include possible data integration and ensure interoperability with other data repositories in related fields. Integrating digital libraries with data grids have been suggested as a way to manage the massive amounts of distributed scientific data that are now being generated (Frey et.al. 2002). The federation of independent digital libraries is being motivated by the desire that individual organization may wish to retain local control over collections, while supports

remote access. The integration of these emerging digital library technologies with federated data grids holds promises to provide the data management infrastructure needed to support universal collaborations (Moore et al. 2005).

4.2 Preservation, Maintenance and Migration Path

Preservation of digital materials has been a concern for quite some time (Lesk 1992; Hedstrom 1997; Rotherberg 1998; Werf-Davelaar 1999; Moore 2003). Studies and workshops have been reported on the subject (Chen 2001b; Ray et al. 2002; NSF 2002). Earlier study from the National Media Laboratory has indicated that the life time of hardware storage (tapes, magnetic disks, CD-ROMS, and other media) is typically less than 5 years (Bogart and John 1996). Digital materials are vulnerable to deterioration and are significantly short lived than traditional format materials (paper, microfilms, etc..) The obsolescence of digital technology manifests itself -- medium itself disappears from market, hardware and equipment are no longer produced, devices and components are obsolete as new technology emerges. Migration of digital materials to new technology is a necessity, but potentially costly, for preservation of a digital repository.

Digital data and documents are inherently software-dependent – besides the hardware environment, they often rely on application software to make them accessible and meaningful. Although not unique to NEES research program, 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 (and often proprietary) software. Certainly, as much as possible, it is desirable that the data be captured in some non-proprietary format and encapsulated in a file or a byte stream as part of a digital object. In reality, it may be necessary to save the software itself with the data. A further complication is found in the heterogeneity of NEES data and documentation, which often include a variety of data types, such as drawings, text, sensor readings, multimedia materials, etc, produced using different platforms and equipment. Preservation of the data in such a variety of formats, software and platforms is not an easy problem to resolve.

Approaches adopted for preservation cannot be separated from the use of the data repository that the present and future users wish to perform. Digital preservation has little value to the research community if it serves only for storage purpose. The only reliable way (and often the only possible way) to access the meaning and functionality of a digital document or the data sets is to run its original software – either the software that created it or some closely related software that understands it (Swade 1998). Maintaining repositories of hardware and software is expensive and not feasible. Various approaches for digital data preservation, such as emulation (Rothenberg 1995, 1998, 2000; Granger 2000); Universal Virtual Computer (Lorie 2001, 2002), infrastructure independence (Moore et.al. 2000a, 2000b) have been proposed. None of the technological approaches have been demonstrated universally useful and it is unlikely that any single approach will work. The choice of preservation technology will depend on digital objects to be preserved, the degree of technical success achieved and on economic and organizational factors (Lord and Macdonald 2003). No matter which approach is undertaken, migration between hardware and software environment requires transforming data from one format to another successively as technologies changes. Each digital application must be migrated forward in time onto new data management systems, simultaneously with the migration of the individual objects onto new media.

Simply put, current methods for preserving digital materials do not fully support the activities needed by a research community. They all involve tradeoffs between what is desirable from the standpoint of functionality, dependability and cost and what is possible and affordable with current technologies and methods. As noted by Hedstroms (1997), "Integration of preservation requirements and methods with access and maintenance systems is essential to fully and efficiently support the processes of migration, regeneration and documentation of the life of digital objects. Planning for preservation must become an integral part of the design and management of digital library and archives. If left as an afterthought, there is little reason to believe that long-term preservation of digital information will be any more affordable than preservation of conventional format has been." Even with excellent file standards, policies, audit trails, etc., files will evolve and new types will arrive over time. Preservation must include the maintenance of the structural characteristics, metadata structure, display, computational and analysis, which rely not only on the mass storage devices but also the software and hardware that may be needed for retrieval and interpretation. Simple migration strategies that involve reformatting of digital materials to a simple standard format are not desirable since they often ignore the structure of documents and relationships embedded in the data sets that are needed in order to meaningfully retrieve the information. 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. Last but not least, any solutions to the preservation problem and migration process must not be labor intensive or error prone. Indeed, preservation and migration strategies could be the most critical decision if a data repository is to endure for the long term.

5 Data Access, Consumption and Collaboration

Data curation cannot be divorced from use. Providing easy access to meaningful and useful data is an important part of a well curated data repository. One purpose of establishing the NEES data repository is to support long-term usage, education and potentially the discovery of new knowledge. The potential users of the NEES repository may include researchers, practitioners, educators, students, product manufacturers, and, potentially, the general public, each may have different interests and needs for access. The initial heaviest users of the archival data will probably be the researchers, investigators, and the students; they should find the easily accessed, well organized, and documented storage of their own data useful. There is a need to identify future users of the data repository and evaluate the requirements, cost and benefits for long term usage of the repository. Tools should be developed to access the data in a systematic manner so that new knowledge and research may be generated from the archived data. Finally, there are many organizations that may have similar and related interests in the earthquake related fields and other scientific communities that are interested in data curation. It will be useful for NEES to collaborate and make partnership with some of these organizations and scientific communities.

5.1 Access Mechanisms

Access to data repositories varies widely, from manual requests to electronic file transfer. With the proliferation of the Internet, web browser has become data access interface of choice (van Veen and Oldroyd 2004). To cope with this trend, academic libraries and digital data centers have dramatically increased their offerings of online resources (Kenney et al. 2002). Even with the web browser, different access mechanisms may be needed depending on the users' experiences and expertise.

Data access mechanism supported by digital archives is different from that of search engines. Search engines typically provide a user with too much noisy information, which may be appropriate for a one-time request for quickly needed information. Digital archives, on the other hand, should provide a user with focused and quality information. It is important to note that the access mechanisms for web search or digital library and archive are neither conflicting nor exclusive, but are complementary in nature. For a digital archive, it is important to pre-define a metadata structure or schema that would be capable of indexing the full range of data types and content. Existing tools that are based on a priori representations that are specific to narrowly defined domains and media types may not be directly suitable for adoption to other domain. Close collaboration between domain experts knowledgeable about the subject and metadata specialists will be necessary for the task. Metadata structure, definitions, vocabulary will evolve as knowledge about access and use increase and it will be an iterative and continuing process to refine the metadata structure and vocabulary over time.

For a user interested in a particular domain, the user often tends to search for focused information specific to a particular area or subject of interests. Therefore, to facilitate users' access to a maze of data sources, browsing and querying capabilities with friendly user interface should be provided (Lassila 1998; Rust 1998; Lacroix et al. 2004). For instance, a user may utilize the interface provided by the digital archive and choose an appropriate way to search the archive. The digital archive returns a reasonable amount of information (with high precision and recall) that the user can readily digest. The returned results are the tertiary scientific data together with high level summary (metadata). Upon request, the relevant primary or secondary research data can be fetched and presented, or ready to be set up for transfer to the user. Depending on users' request, a proactive mechanism could also be useful to inform users about related and relevant information and to deliver by pushing the information to the users (Ye and Fischer 2002). One distinguish feature that the NEES data access may differ from typical digital library search is the amount of data that may return to the users, even for a simple request. The data may be in a wide variety of formats, from text files to images and video. Furthermore, in certain cases, application programs that operate on the data may be needed to allow the users to visualize and to understand the information retrieved. Thus, an "information delivery" mechanism where the system can deliver to the user the information in appropriate media may be necessary.

Among the long term goal of NEES research is data reuse and knowledge discovery. Information retrieval and knowledge discovery process is highly dependent on universal homogeneous access to heterogeneous information. To support logical inference and reasoning, it is necessary to support relationships and rules imposed as constraints when accessing the data. The mechanisms may vary from extracting parts of the electronic materials, with full text annotation, converting parts of the information for viewing to download, and to support applications. To fully support research investigation and knowledge discovery, the desirable features may include:

- the ability to identify relationships between digital objects, using the stored data for scientific discovery.
- the ability to query across multiple data repositories to identify data sets of interest.
- the ability to read data from a remote site for incorporation and use within an application.
- the ability to filter a data set for transmission over the network.
- the ability to add data sets to collections for use by other researchers.
- the ability to use data in scientific simulations, data mining, and creation of new data collections.

Such features will provide the direct supports needed by an end user to further conduct research using a data repository.

Finally, access mechanisms may also be needed to support inexperienced users by providing alternative approaches and delivery mechanisms to facilitate users' search. In terms of accessing the data objects based on users' roles, the NEES repository needs to present the data to different audiences: K-12 students, general public, researchers, and data creators. Access entails the retrieval of relevant information and makes it available to the users in a suitable way. In summary, while it is possible to adopt a single access point to the repository, user interfaces and information delivery mechanisms will likely be different for different users and applications.

5.2 Collaboration

There are many ongoing works in the digital library and scientific communities that are building data repositories. It would be beneficial for NEES to collaborate and to establish partnership with some of these organizations. It will be useful if NEES data sources can be integrated and/or interoperable with complementary data sources from other communities in earthquake engineering and seismology. 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. Linkage to other related information centers, such as the the Consortium of Universities for Research in Earthquake Engineering (CUREE) Earthquake Engineering Research Institute (EERI), The Incorporated Research Instutions for Seismology (IRIS), United States Geological Survey (USGS) and others, could be of interests to a broad audience. Partnership building with other international earthquake engineering centers, such as the National Center for Research on Earthquake Engineering (NCREE) in Taiwan, Hyogo Earthquake Engineering Reserch Center (E-Defense) in Japan, Korea Construction Engineering Development Program (KOCED) in South Korea and others, who have similar research program, could help expand the data content and collections to enrich the NEES data repository. Metadata development effort can also be benefited from other similar efforts such as the web-based Electronic Encyclopedia of Earthquakes (E^3) (see http://sceccore.usc.edu/e3/index.php), SensorML and data standards for The Consortium of Organizations for Strong-Motion Observation Systems (COSMOS).

In addition to earthquake related efforts, there have been a number of collaborative science projects that have similar objectives as NEES. These include, for examples, the National Virtual Observatory (NVO) (see http://www.us-vo.org), the CyberInfrastructure for the Geosciences (GEON) (see http://www.geongrid.org), Science Environment for Ecological Knowledge (SEEK) (see http://seek.ecoinformatics.org), Sloan Digital Sky Survey (Skyserver) (see http://cas.sdss.org/), Scripps Institute of Oceanography (SIOExplorer) (see http://nsdl.sdsc.edu/), Inter-University Consortium for Political and Social research (ICPSR) (see http://www.icpsr.umich.edu/), etc.. It should be noted that none of the current collaborative science projects (that the authors are aware of) are as diverse as the collaborative experimental and simulation research as NEES. Nevertheless, lessons can be learnt from many of these ongoing repository development projects.

There are also many organizations that have extensive ongoing efforts in digital library, archiving and preservation. These include, for examples, LOCKSS (see http://lockss.stanford.edu), DSPACE (see http://www.dspace.org), NEDLIB (see http://www.kb.nl), Fedora (see http://www.fedora.info), CODA

(see http://library.caltech.edu/digital), LOC (see http://www.digitalpreservation.gov), etc.. These ongoing efforts could provide many insights towards metadata development, state of practice in digital preservation and archiving.

Last but not least, the emergence of semantic web and grid computing technologies will have significant impact to collaborative science research (Foster and Kesselman 1999; Berners-Lee et al. 2001; Berman et al 2003). Examples of these efforts include the Open Science Grid (OSG) (see http://www.opensciencegrid.org), and the Enabling Grids for EscienceE (EGEE) (see http://public.eu-egee.org).

It should be noted that the works cited above are not meant to be exhaustive, but are merely representative examples of the many on-going activities that may be of interests to the NEES efforts.

6 Other Administrative and Policy Issues

In addition to the topics discussed in the previous sections, there are a number of administrative issues that need to be concerned with in order to put together a well-curated data archive.

- *Data Quality and Control*: A standard data verification process is needed to guarantee the integrity of data. Creation date, versions and digital signature may be needed to verify the data sets (raw, original versus derived, etc.). Quality control, including the status of the data ingested (unedited, edited), is important to ensure that the data in the repository can be trusted. Quality assurance of data and practice has been identified as one of the most important tasks by the NEES's Data Sharing and Archiving Committee (NEES-DSAC 2004).

- *Data Ownership, Rights and Security*: Related to a public data repository is the policy issue that has to do with the rights and authority over the data (see, for example, the Cedars Project 2002). Who has the right to make changes to the data? How strict should the "permission to use" policy be enforced? Should the researchers be allowed to withhold certain restricted data sets? Will researchers be allowed to modify (or correct) the datasets (and for how long) after the submission of "final" project report? What rights does the archive have? What rights do various user groups have? Who "own" the content in the repository? What rights has the owner retained? How will the access mechanism interact with the archive's metadata to ensure that these rights are managed properly? Rights management includes providing or restricting access as appropriate, and changing the access rights as the material's copyright and security level changes. Besides right management, security and version control are also important to digital archives. Secure access to data repository must be enforced so that the digital objects saved in the repository will not be tampered with.

- *Audit Trail*: Data management should also provide support of an audit trail over the lifecycle of the data. The audit trail information should be tied closely with metadata 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 important for building confidence on the data. Ability to establish authenticity and integrity of the archived data sets are important as part of the data management scheme (Lynch 1994; Lord and McDonald 2003).

The workflow process and management policies for data curation must take into consideration the issues mentioned above.

7 Summary and Discussion

This report has discussed a few selected issues that may deem relevant to the NEES data repository and curation effort. Specifically, the discussion follows the basic functional modules – data ingestion, management and preservation and access -- as described in the Open Archive Information System (OAIS) framework.

In the data production (generation/ingestion) area, the report discusses some of data selection, metadata and ingestion issues that may worth further examination. Analogous to the relationship between a book author and the library, the data ingestion process, repository design and the data curation effort should be sufficiently robust such that there will not be any unnecessary burdens imposed onto the researchers (the data producers). The types of NEES data that need short term storage and those that worth long term archival and preservation value should be defined. Metadata structure, definitions and elements that are suitable for NEES experiments should be established. Taxonomy and vocabulary should be captured to support access functions. Data and metadata ingestion tools need to be developed to assist researchers and equipment sites to "load" project information (including experimental and simulation data) into temporary and permanent repository. From the data ingestion and management perspectives, it may be worthwhile to differentiate the physical storage and logical organization (schema) of the data collection as well as the management of data repository and the metadata for cataloging. Mappings can be added to reference metadata information with data entities in the data repository to facilitate discovery and browsing. Last but not least, positive incentives should be provided to encourage researchers to submit as much valuable data as possible to the repository. As Richard Feynman noted in his Nobel Lecture (1966), "We have a habit in writing articles published in scientific journals to make the work as finished as possible, to cover up all the tracks, to not worry about the blind alleys or describe how you had wrong idea first, and so on. So there isn't any place to publish, in a dignified manner, what you actually did in order to get to do the work." Advances in storage and network technologies continue to make storing and managing scientific data economical. With appropriate incentives, a digital, scientific data repository, with the capability to capture all the data and information regarding a research project, could become a valuable instrument to capture all phases (from the pre-design to the completion) and the different (positive and negative) aspects of the project.

Among the key issues in data management are the repository design, maintenance strategies and preservation. Data management and maintenance include backup strategies, data quality control, audit trail, version management and other technical issues. These technical issues could be resolved by following current and future digital library practice. Another dimension of the data repository design is to develop a plan that will support interoperability and integration with other data repositories in earthquake engineering and related fields. Perhaps the most difficult long term problem is the issue of digital data preservation and migration. This problem has been recognized as one of the urgent technical hurdles faced by the digital library and data archiving. In December 2000, the US Congress passed a legislation to establish and has asked the Library of Congress (LoC) to lead an effort to develop a National Digital Information and Infrastructure and Preservation Program (NDIIPP) (http://www.digitalpreservation.gov/). While the focus of NDIIPP is on ensuring preservation of historically significant digital content, much of the activities can have significant ramifications to

preservation and management of NEES data. In 2004, LoC and the National Science Foundation (NSF) has initiated the Digital Archiving and Long Term Preservation (DIGARCH) research program to further address this important problem. A number of current projects are related to the management and preservation of scientific data. Decisions about preserving information should consider data migration issue and the long term cost. Technologies adopted need to become affordable. A dialogue among preservationists, library scientists and scientific researchers is important to define an appropriate agenda for the digital archiving of scientific data such as NEES's (Hedstrom 1997).

Data ingestion, management and preservation cannot be separated from use. First and foremost is to understand the audience that the data repository is intended to serve in the short term and in the long term. What data to be ingested and preserved and what mechanisms to access the data depend highly on the targeted audience. From the research perspective, besides the browsing and querying capabilities, metadata structure and retrieval tools should be developed to support data reuse, mining and knowledge discovery. Furthermore, interoperability with other data repositories in related fields will be desirable. Last but not least, collaborations and partnerships with other organizations in earthquake engineering, seismology, digital library and preservation that have interests in data collection activities will be beneficial to NEES and will prevent duplicated efforts from reinventing the wheel while focusing the resources on the needs and the mission of NEES. Relevant technologies should be assessed and evaluated to establish a technical strategy. Long term commitments and "business" model are needed to sustain and facilitate optimal usage of the data in the repository.

In summary, data curation strategies and access tools are very important part of the NEES program. A data repository is only as good and useful as what can be done with the stored data to further enhance the science and the education. Strategic plans, including goals, administrative policies, technologies, costs and benefits for the short (< 5 years), medium (5-10 years) and long term (> 10 years) should be established. Curating a collection is an arduous task that requires deep knowledge of the scientific discipline as well as information science and tools (Atkins et.al. 2003). Researchers must be willing to spend efforts to ingest data for archival. Librarian should consult with earthquake engineering and related communities to establish appropriate metadata structure and work closely with IT team to integrate with the data repository. Technologies need to be acquired or developed to support data ingestion, management and preservation, and access. The management team should develop a plan to define the intention and scope and to carefully evaluate the cost and the benefits. Needless to say, the importance of data curation technologies goes beyond the earthquake engineering domain. The scientific enterprise and "cyberinfrastructures" will be benefited from the experience gained from the NEES's data repository and curation efforts.

References

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

Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., Messina, P., Messerschmitt, D. G., Ostriker, J. P. and Wright, M. H. (2003). *Revolutionizing Science and Engineering Through Cyberinfrastructure*. Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastrure, (available at http://www.communitytechnology.org/nsf_ci_report/).

Baca, M. (ed.) (2000), *Introduction to Metadata - Pathways to Digital Information* (available at http://www.getty.edu/research/institute/standards/_intrometadata.)

Baldonado, M., Chang, C.-C. K., Gravano, L. and Paepcke, A. (1997a). "Metadata for Digital Libraries: Architecture and Design Rationale." *Proceedings of 2nd ACM International Conference on Digital Libraries (DL'97)*, Philadelphia, PA.

Baldonado, M., Chang, C.-C. K., Gravano, L. and Paepcke, A. (1997b). "The Stanford Digital Library Metadata Architecture." I*nternational Journal of Digital Library*, 1(2) pp. 108-121.

Bass, M.J., Stuve, D., Tansley, R., Branschofsky, M., Breton, P., Carmichael, P., Cattey, B., Chudnov, D., Ng, J. (2002). *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)

Berman, F., Fox, G. and Hey, A. J. G. (2003). *Grid Computing: Making the Global Infrastructure a Reality*, Wiley.

Berners-Lee, T., Hendler, J. and Lassila, O. (2001). "The Semantic Web." *Scientific American*, 284(5) pp. 34-43.

Berry, D., Buneman, P., Wilde, M. and Ioannidis, Y. (2003). *Workshop on Data Derivation and Provenance*. Edinburgh, Scotland, (available at http://www.nesc.ac.uk/esi/events/304/).

Bogart, V. and John, W. C. (1996). "Long-Term Preservation of Digital Materials." *Proceedings of the National Preservation Office Conference on Preservation and Digitisation: Principles, Practice and Policies*, University of York, England.

Botts, M. (ed.) (2002). *Sensor Model Language (SensorML) for In-situ and Remote Sensors*. OpenGIS Interoperability Program Report, OGC 02-026, Open GIS Consortium Inc, (available at http://vast.uah.edu/SensorML/SensorML_04-019_1.0_beta.pdf).

Buneman, P. and Foster, I. (2002). *Workshop on Data Derivation and Provenance*. Chicago, IL, (available at http://www-fp.mcs.anl.gov/~foster/provenance/).

Buneman, P., Khanna, S., Tajima, K. and Tan, W.-C. (2002). "Archiving Scientific Data." *Proceedings of 2002 ACM SIGMOD International Conference on Management of Data*, Madison, Wisconsin, pp. 1-12.

Chen, C. (2001a). "Global Digital Library Development." *Knowledge-based Data Management for Digital Libraries*, Tsinghua University Press, Beijing, China, pp. 197-204.

Chen, P.P.-S. (1976). "The Entity Relationship Model – Toward a Unified View of Data," *ACM Transactions on Database Systems*, 1(1):9-36.

Chen, S.S. (2001b), "The Paradox of Digital Preservation," *IEEE Computer*, March, pp. 24-28.

Codd, E.F. (1982) "Relational Database: A Practical Foundation for Productivity," *Communications of the ACM*. Vol. 25, no. 2, pp. 109-117.

Conway, P. (1990). "Archival Preservation in a Nationwide Context." *American Archivist*, 53(2):204–222.

Cooper, B., Crespo, A. and Garcia-Molina, H. (2000). "Implementing a Reliable Digital Object Archive." P*roceedings of Fourth European Conference on Research and Development in Digital Libraries (ECDL)*, Lisbon, Portugal.

Cooper, B. and Garcia-Molina, H. (2001). "Creating Trading Networks of Digital Archives." *Proceedings of the First ACM/IEEE Joint Conference on Digital Libraries*, Roanoke, Virginia.

Date, C.J. (2004), *An Introduction to Database Systems*, 8th Edition, Addison Wesley.

DCMI. (2005). The Dublin Core Metadata Initiative. (see http://purl.oclc.org/).

EML. (2005). Ecological Metadata Language (EML) Specification. Technical Report No. (available at http://knb.ecoinformatics.org/software/eml/eml-2.0.1/index.html).

Fenves, G.L. and McKenna, F. (2004), *Data Model for Simulation*, Technical Report NEESgrid-2004-46. (available at http://it.nees.org/documentation/pdf/TR_2004_46.pdf)

Foster, I. and Kesselman, C. (1999). *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann, San Fransisco, CA.

Frey, J. G., Roure, D. D. and Carr, L. A. (2002). "Publication At Source: Scientific Communication from a Publication Web to a Data Grid." *Proceedings of Euroweb 2002 Conference, The Web and the GRID: from e-science to e-business*, Oxford, UK.

Garas, F. K. and Hunter, I. (1998) "CIMSteel (Computer Integrated Manufacturing in Constructional Steelwork) - Delivering the Promise," *Structural Engineering*, 76(3):43-45.

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

Goldberg, A. V. and Yianilos, P. N. (1998). "Towards an Archival Intermemory." *Proceedings of IEEE International Conference of Advances in Digital Libraries (ADL'98)*, Los Alamitos, CA, 147-156.

Granger, S. (2000) "Emulation as a Digital Preservation Strategy," *D-Lib Magazine*, Volume 6, No. 10. (available at http://www.dlib.org/dlib/october00/granger/10granger.html).

Gray, J., Liu, D. T., Nieto-Santisteban, M. A., Szalay, A. S., Heber, G. and DeWitt, D. (2005). *Scientific Data Management in the Coming Decade*. Microsoft Corporation, MSR-TR-2005-10, (available at ftp://ftp.research.microsoft.com/pub/tr/TR-2005-10.pdf).

Gray, J., Szalay, A. S., Thakar, A. R., Stoughton, C. and VandenBerg, J. (2002). "Online Scientific Data Curation, Publication, and Archiving." *Proceedings of SPIE Astronomy Telescopes and Instruments*, Waikoloa, Hawaii.

Heath, B.P., McArthur, D.J., McClellang, M.K., and Vetter, R.J. (2005), "Metadata Lessons from the iLumina Digital Library," *Communications of the ACM*, 48(7):68-74.

Hedstrom, M. (1997). "Digital Preservation: A Time Bomb for Digital Libraries," *Computers and the Humanities*, 31(3):189-202.

Hedstrom, M. (2005). "Incentives for Data Producers to Create 'Archive-Ready' Data Sets." DIGARCH PIs Meeting, Atlanta, GA http://diggov.org/library/library/dgo2005/digarch/hedstrom.pdf.

Hillmann, D. (2005). "NSDL Metadata Primer." (available at http://metamanagement.comm.nsdlib.org/overview2.html)

Hodge, G. M. (2000). "Best Practices for Digital Archiving: An Information Life Cycle Approach." *D-Lib Magazine*, 6(1) (available at http://www.dlib.org/dlib/january00/01hodge.html).

Hovy E. (2003). "Using an Ontology to Simplify Data Access," *Communications of the ACM*, 46(1):47-49.

IEEE (2002), *Standards for Information Technology – Education and Training Systems – Learning Objects and Metadata (LOM).* IEEE Standards 1484.12.1-2002. (available at http://Itsc.ieee.org/wg12/).

IAI (1997). *Industry Foundation Classes*, Specification Volumes 1-4, International Alliance for Interoperability, Washington, D.C..

ISO (1994). *Product Data Representation and Exchange, Part 1: Overview and Fundamental Principles*, No. 10303-1, International Organization for Standardization, 1994.

ISO (2003), *Information and Documentation – The Dublin Core Metadata Element Set, ISO 15836. (available at www.niso.org/international/SC4/n515.pdf)*

Jaja, J. (2005). "Robust Technologies for Automated Ingestion and Long Term Preservation of Digital Information." DIGARCH PIs Meeting, Atlanta, GA http://diggov.org/library/library/dgo2005/digarch/jaja.pdf.

Jones, M. (2001). "The CEDARS Project Website." *D-Lib Magazine*, 7(12) (available at http://www.dlib.org/dlib/december01/12inbrief.html).

Kenney, A. R., McGovern, N. Y., Botticelli, P., Entlich, R., Lagoze, C. and Payette, S. (2002). "Preservation Risk Management for Web Resources: Virtual Remote Control in Cornell's Project Prism." *D-Lib Magazine*, 8(1) (available at http://www.dlib.org/dlib/january02/kenney/01kenney.html).

Kim, Y.S. (2004). *SAC Metadata for NEESgrid*, Technical Report NEESgrid-2004-09. (available at http://it.nees.org/documentation/pdf/TR_2004_09.pdf)

Kubiatowicz, J., Bindel, D., Chen, Y., Czerwinski, S., Eaton, P., Geels, D., Gummadi, R., Rhea, S., Weatherspoon, H., Weimer, W., Wells, C. and Zhao, B. (2000). "OceanStore: An Architecture for Global-Scale Persistent Storage." *Proceedings of the Ninth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2000)*, Cambridge, MA.

Kutter, B.L., Wilson, D.W., Bardet, J.P. (2002), *Metadata Structure for Geotechnical Physical Models (and Simulations?)*, Technical Report. (available at http://it.nees.org/documentation/pdf/mtml.pdf).

Lacroix, Z., Parekh, K., Raschid, L. and Vidal, M. E. (2004). "Navigating through the Biological Maze." *Proceedings of IEEE Computational Systems Bioinformatics Conference (CSB'2004)*, Stanford, CA.

Lagoze, C., Payette, S., Shin, E. and Wilper, C. (2005). "Fedora: An Architecture for Complex Objects and their Relationships," forthcoming in *Journal of Digital Libraries, Special Issue on Complex Objects*, Springer 2005.
(available at http://www.arxiv.org/abs/cs.DL/0501012)

Lassila, O. (1998). "Web Metadata: A Matter of Semantic." *IEEE Internet Computing*, 2(4) pp. 30-37.

Lau, G. T. (2004) *A Comparative Analysis Framework for Semi-Structured Documents, with Applications to Government Regulations*, Ph.D. Thesis, Department of Civil and Environmental Engineering, Stanford University, Stanford, CA, August 2004.

Lau, G.T., Law, K.H. and Wiederhold, G. (2005a). "A Relatedness Analysis Tool for Comparing Drafted Regulations and the Associated Public Comments," *Journal of Law and Policy for the Information Society* , 1(1):95-110.

Lau, G.T., Law, K.H. and Wiederhold, G. (2005b). "A Relatedness Analysis of Government Reulgations Using Domain Knowledge and Structural Organization," (submitted) *Information Retrieval.*

Lesk, M. (1992). *Preservation of New Technology: A Report of the Technology Assessment Advisory Committee to the Commission on Preservation and Access.* Washington, D.C., Commission on Preservation and Access.

Liskov, B., Ghemawat, S., Gruber, R., Johnson, P. and Shrira, L. (1991). "Replication in the Harp File System." *Proceedings of Thirteenth ACM Symposium on Operating Systems Principles*, Pacific Grove, CA, pp. 226-238.

Lord, P. and Macdonald, A. (2003). *e-Science Curation Report. Data Curation for e-Science in the UK: an Audit to Establish Requirements for Future Curation and Provision*. The Digital Archiving Consultancy Limited, Twickenham, UK, (available at http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf).

Lorie, R. (2001). "Long Term Preservation of Digital Information." *Proceedings of First ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'01)*, Roanoke, Virginia.

Lorie, R. (2002). *The UVC: a Method for Preserving Digital Documents: Proof of Concept*. IBM/KB Long-Term Preservation Studies Report Series 4, (available at http://www.kb.nl/hrd/dd/dd_onderzoek/reports/4-uvc.pdf).

Lynch, C. (1994) "The Integrity of Digital Information: Mechanics and Definitional Issues," *Journal of the American Society for Information Science*, 45(10):737-744.

Lynch, C. (1999). "Canonicalization: A Fundamental Tool to Facilitate Preservation and Management of Digital Information," *D-Lib Magazine*, 5(9). (available at http://www.dlib.org/dlib/september99/09lynch.html)

Moore, R. (2000). "Data Management Systems for Scientific Applications." *Proceedings of The Architecture of Scientific Software*, Ottawa, Canada, pp. 273-284.

Moore, R. (2003). *Preservation of Data*. SDSC Technical Report 2003-06, San Diego, CA, (available at http:// www.sdsc.edu/dice/Pubs/data-preservation.doc).

Moore, R., Baru, C., Rajasekar, A., Ludaescher, B., Marciano, R., Wan, M., Schroeder, W. and Gupta, A. (2000a). "Collection-Based Persistent Digital Archives - Part 1." *D-Lib Magazine*, 6(3) (available at http://www.dlib.org/dlib/march00/moore/03moore-pt1.html).

Moore, R., Baru, C., Rajasekar, A., Ludaescher, B., Marciano, R., Wan, M., Schroeder, W. and Gupta, A. (2000b). "Collection-Based Persistent Digital Archives - Part 2." *D-Lib Magazine*, 6(4) (available at http://www.dlib.org/dlib/april00/moore/04moore-pt2.html).

Moore, R. W., Rajasekar, A. and Wan, M. (2005). "Data Grids, Digital Libraries and Persistent Archives: An Integrated Approach to Publishing, Sharing and Archiving Data." *Proceedings of the IEEE*, 93(3) pp. 578-588.

Murray-Rust, P., Rzepa, H. S. and Wright, M. (2001). "Development of Chemical Markup Language (CML) as a System for Handling Complex Chemical Content." *New Journal of Chemistry*, 25(4) pp. 618-634.

NEES-DSAC. (2004). *NEES Consortium, Inc. (NEESinc) Data Sharing and Archiving Policies and Guidelines*. Network for Earthquake Engineering Simulations Report.

NOST. (1999). *Definition of the Flexible Image Transport System (FITS)*. NASA/Science Office of Standards and Technology, NOST 100-2.0, (available at http://archive.stsci.edu/fits/fits_standard/).

NSF. (2002). *It's About Time: Research Challenges in Digital Archiving and Long-Term Preservation*. Workshop on Research Challenges in Digital Archiving and Long-Term Preservatiopn, Final Report, (available at http://www.digitalpreservation.gov/repor/NSF_LC_Final_Report.pdf).

OAI, (2005), "Open Archive Initiative," (see http://www.openarchives.org/.

OAIS. (2002). Reference Model for an Open Archival Information Systems. Technical Report No. CCSDS 650.0B-1, Consultative Committee for Space Data Systems,, (available at http://ssdoo.gsfc.nasa.gov/nost/isoas/).

Oregon State University (OSU) and Network Alliance for Computational Science and Engineering (2003). *NEES Database and Metadata Structure*, Version 1.3, white paper, Network for Earthquake Engineering Simulation.

OWL (2004). *Web Ontology Language Reference* (available at http://www.w3.org/TR/owl-ref/)

Peng, J. and Law, K. H. (2004a). *A Brief Review of Data Models for NEESgrid*. Report NEESgrid-2004-01, (available at http://it.nees.org/documentation/pdf/TR_2004_01.pdf).

Peng, J. and Law, K. H. (2004b). *Reference NEESgrid Data Model*. Report NEESgrid-2004-40, (available at http://it.nees.org/documentation/pdf/TR-2004-40.pdf).

Peng, J. and Law, K. H.(2005) *Data Retrieval Tools: Software Design Specification*, NEESit Technical Report, 2005.

Rajasekar, A. and Moore, R. (2001). "Data and Metadata Collections for Scientific Applications." *Proceedings of the 9th International Conference on High-Performance Computing and Networking*, Amsterdam, The Netherlands, 72 - 80.

Ray, J., Reich, V., Dale, R., Underwood, W., Moore, R. and McCray, A.T. (2002), "Panel on Digital Preservation," *Joint Conference on Digital Libraries (JCDL)*, Portland, Oregon.

Rothenberg, J. (1995). "Ensuring the Longevity of Digital Documents." *Scientific American*, 272(1) pp. 42-47.

Rothenberg, J. (1998). *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation*. Council on Library and Information Resources, (available at http://www.clir.org/pubs/reports/rothenberg/contents.html).

Rothenberg, J. (2000). *An Experiment in Using Emulation to Preserve Digital Publications*. NEDLIB Report, The Koninklijke Bibliotheek and RAND-Europe, (available at http://www.kb.nl/coop/nedlib/results/emulationpreservationreport.pdf).

Rust, G. (1998). "Metadata: The Right Approach. An Integrated Model for Descriptive and Rights Metadata in E-commerce" *D-Lib Magazine*, July/August (available at http://www.dlib.org/dlib/july98/rust/07rust.html).

Salton, G. and McGill, M. (1983). *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, NY.

Sall, K. B. and Ames, T. J. (1999). *Using XML and Java for Astronomical Instrument Control*. NASA, (available at http://isd.gsfc.nasa.gov/Papers/DOC/sall.html).

Sharon, T. and Frank, A. J. (2000). "Digital Libraries on the Internet." P*roceedings of 66th IFLA Council and General Conference*, Jerusalem, Israel.

Swade, D. (1998). "Preserving Software in an Object-Centred Culture." *History and Electronic Artifacts*, E. Higgs, ed., Oxford: Clarendon Press, pp. 195–206.

Swift, J., Eng, J., Bardet, J.-P., Lui, F., Mokarram, N., and Pekcan, G. (2004), *Using the NEES Reference Data Model and the NEES Metadata Browser for Centrifuge Experiments*, Technical Report, Version 1.1, (available at http://it.nees.org/documentation/pdf/TR-2004-50.pdf)

Teorey, T.J. (1998). *Database Modeling & Design*, Third Edition Morgan Kaufmann.

The Cedars Project. (2002). *Cedars Guide To: Intellectual Property Rights*. (available at http://www.leeds.ac.uk/cedars/guideto/ipr/guidetoipr.pdf).

Ullman, J.D. (1989), *Principles of Database and Knowledge-Based Systems*, Vols. 1 & 2, Computer Science Press.

van Veen, T. and Oldroyd, B. (2004). "Search and Retrieval in the European Library: A New Approach." *D-Lib Magazine*, 10(2) (available at http://www.dlib.org/dlib/february04/vanveen/02vanveen.html).

Warnock, T. (2005). *Central Repository Design Specification*. NEESit Report TR-2005-006, (available at http://it.nees.org/documentation/pdf/TR-2005-006.pdf.)

Werf-Davelaar, T. v. d. (1999). "Long-term Preservation of Electronic Publications: The NEDLIB project." D-Lib Magazine, 5(9) (available at http://www.dlib.org/dlib/september99/vanderwerf/09vanderwerf.html).

Ye, Y. and Fischer, G. (2002). "Supporting reuse by delivering task-relevant and personalized information." *Proceedings of the Twenty-Fourth International Conference on Software Engineering (ICSE)*, Orlando, FL, pp. 513-523.