The NEEScentral Data Repository: A Framework for Data Collaboration in Earthquake Engineering

L. Van Den Einde[1], K. Fowler[1], S. Krishnan[1], J. Rowley[1], K. Bhatia[1], C. Baru[1], A. Elgamal[2]

[1] *NEES Cyberinfrastructure Center, San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA,USA*

[2] *Professor, Dept of Structural Engineering University of California, San Diego, La Jolla, CA,USA*

Email: lellivde@sdsc.edu, kfowler@sdsc.edu, jrowley@sdsc.edu, sriram@sdsc.edu, baru@sdsc.edu, elgamal@ucsd.edu

ABSTRACT :

The NEES Cyberinfrastructure Center (NEESit) has developed a central data repository (NEEScentral) to capture important reservoirs of data and expose them to the community where their useful lifetime can be extended. The NEEScentral Data Repository provides a centralized location for researchers to securely organize, store, and share data and metadata in a nonproprietary format that can be used in data manipulation and visualization tools. NEEScentral supports the *NEES Data Model*, which defines the structure by which NEES experiment data is organized. Using NEEScentral, users can upload, view, and download data and metadata. This paper provides an overview of the system architecture for the NEEScentral Data Repository, and describes features in support of earthquake engineering research including the ability to upload data, perform novel searches, create and print customizable reports of the data and metadata, curate and publish data sets, download data sets in formats that allow for ingestion into community developed visualization or data processing programs, and export project and experiment data and metadata from NEEScentral. The paper highlights the benefits of putting research data into the repository.

KEYWORDS: Data, Repository, NEES, Data Repository, Data Model, NEES, Collaboration

1. INTRODUCTION

The George E. Brown Jr. Network for Earthquake Engineering Simulation (NEES) is a project funded by the National Science Foundation (NSF) to provide a widely accessible, state-of-the-art science and engineering infrastructure through experimental testing facilities and the application of advanced computers, networking, and software. NEES integrates 15 facilities that provide equipment for NEES research and shared use projects through a unique infrastructure managed by the NEES Cyberinfrastructure Center (NEESit), a multi-institutional organization supporting the IT needs of the earthquake engineering community, including researchers, practitioners, educators, students, and information technology specialists [1]

In many cases in the past, experimental and numerical data generated by earthquake engineering research experiments have been inadvertently lost after the completion of a project. In other cases the preserved data may be distributed geographically across the globe and syntactically across many formats. As a result, researchers wishing to examine results from multiple data sets must collect the data from each project manually, and often must reorganize them before they can be used. Even data that have been preserved from past projects by a single laboratory or equipment site facility are often lacking any standard organization or the metadata required to combine the information into a useful data set since, typically, once the graduate student or post-doctoral researcher leaves the institution, the specific details about the experimental programs are lost. Long-term preservation of properly curated data in an accessible format is essential in order to ensure that future research can benefit from current earthquake engineering experiments, shortening the gap between research and practice. The ultimate goal for NEESit is to provide access to different types of experimental and computational data to promote and facilitate collaborative and interactive processes required to address the complex nature of seismic events and their physical and societal impacts, ultimately reducing vulnerabilities from earthquakes.

Another important goal of the IT infrastructure is to promote collaboration between geographically distributed

participants. To support this objective, collaborators must be able to easily share documents and organize effective online meetings where they can share ideas and information on planning and performing experiments and analyzing results. The software and services supported by NEESit provides the IT infrastructure to 1) store and share data in a centralized data repository, 2) enable collaboration between distributed researchers, 3) facilitate remote participation in experiments and visualization capabilities, 4) support simulation, including hybrid testing, and 5) securely enable authorized access to data and resources.

A centralized data repository with formalized organizational standards would help preserve this important (and expensive) data and provide the community with a single place for locating past historical data and as well as sharing new data. By providing access to different types of physical and numerical experimental data, previous topic-oriented individual research efforts can be transformed to collaborative and interactive processes required to address the complex nature of seismic events and their physical and societal impacts which will ultimately reduce vulnerabilities from earthquakes [2]. With access to an entire community's data, many new opportunities arise, including the ability to easily compare parameters from parameter studies using past data sets, to validate test results reported by fellow researchers and to construct more accurate simulation models. By reducing the reliance on expensive and time consuming physical experimentation techniques, these more accurate numerical models stand to dramatically improve the amount of research that could be conducted at any given time. These improvements in the quantity and accuracy of earthquake engineering research will provide the foundation for designing buildings and lifelines to reduce losses from earthquakes and could help shorten the time required for research results to be adopted by practitioners into design codes.

This paper provides an overview of the system architecture for the NEEScentral Data Repository, and describes features in support of end-to-end earthquake engineering research such as the ability to easily upload data, perform novel searches, create and print a customizable report of the data and metadata, curate and publish data sets, download data sets in formats that allow for easy ingestion into community developed visualization or data processing programs, and easy export of project and experiment data and metadata. Furthermore, the paper highlights the goals of NEEScentral and benefits for putting research data into the repository.

2. THE NEES CENTRAL DATA REPOSITORY (NEEScentral)

2.1. NEEScentral Architecture

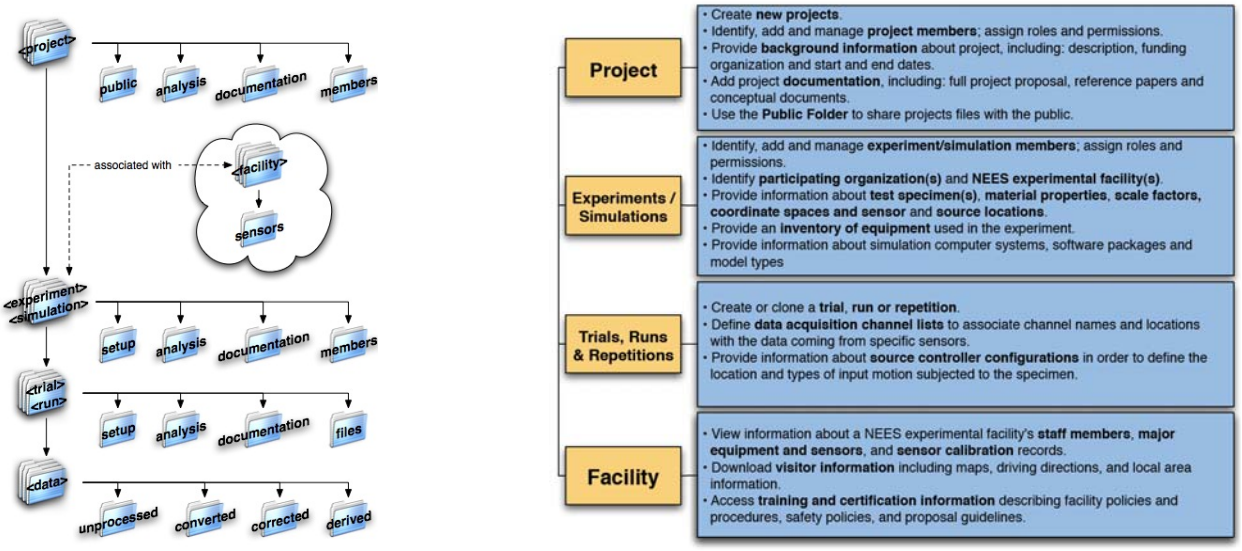
NEEScentral is open source software that was originally built on a standard LAMP software stack to run the dynamic Web server. LAMP stands for Linux (operating system), Apache (web server), MySQL (database management system), and PHP (programming language) [3]. In June 2008, the database back-end was migrated to Oracle version 10g, to take advantage of its built-in capabilities for constraint management, reliability, and scalability. Furthermore, the database triggers and the ability to program procedures (PL/SQL) in Oracle are very useful for application development and to enhance performance. .

To ensure the safety, availability and security of the data entered into the NEEScentral repository, the database housed at the San Diego Supercomputer Center (SDSC) is replicated to the UNAVCO facility in Boulder Colorado using Oracle's *DataGuard* software for data replication. The *DataGuard* software uses "log-based" replication—it uses database transactions logged in the primary database at SDSC and applies them to the offsite to the standby database. Thus, in the event of any failure and system outage, the system can switch to the standby database and operations can continue as normal. Also, in the event of a media failure or some other form of data corruption in the primary database, the data can be recovered using the secondary database. NEEScentral employs a backup strategy composed of full database backup archives, continuous logs and backing up of the transaction logs.

3. THE NEES DATA MODEL

Data in the repository is organized in a conceptual hierarchy as shown in Figure 1. The hierarchy provides a single, high-level, standardized model for storing data that is universal to all disciplines in the NEES community. More detailed information about the NEESit data model and the specific metadata collected can be found in [4]. The organizational hierarchy is based on four directory levels: *project*, *experiment* or *simulation*, *trial* or *run*, and *data*. At the coarsest level of the hierarchy is the project directory, which provides a general container for

secure data sharing within a collaborative team. The second level provides extensive facilities for inputting metadata about the setup of an experiment or simulation, including information about input motions, coordinate spaces, sensor location plans, and scale factors. Experiments and simulations contain one or more trials or runs that define further changes to configuration parameters. At the lowest level of the hierarchy are directories for archiving the data associated with each experiment [5].



a) Hierarchy Structure b) Functions at Each Level

Figure 1: NEEScentral Data Model Hierarchy

As previously mentioned, an important focus of NEESit is to provide the so-called "end-to-end" data usage capability, from data upload, to discovery, and access, by structuring and organizing the data from experiments and simulations with adequate metadata to effectively describe the research. This structuring of the data allows for novel searches to be conducted to easily discover data sets of interest. The challenge is striking a balance between structuring the data necessary to provide insightful queries and make the data sets understandable and reusable by the broader community versus overburdening the data and metadata ingestion requirements so that researchers are reluctant to spend the time to upload their data/metadata. The publication of data sets to the broader community ultimately facilitates use of the research results by practitioners and will lead to the improvement of the state-of-practice through the evolution of design codes.

Once basic data and metadata describing the simulation or experiment are uploaded to the NEES repository, they undergo several curation steps. From the NEEScentral web application, a user can access/download the curated data and import them into analysis tools such as Matlab for data mining, as well as other rendering and visualization tools (EditorM, RDV, and community developed tools as described in [7]). Any processed data can be easily ingested into NEEScentral and made available for others to use.

4. FEATURES OF THE NEEScentral APPLICATION

4.1. Data and Metadata Upload

4.1.1. NEEScentral User Interface

Several methods are available for upload of data and metadata into NEEScentral. One is using a forms-based Web interface, where different forms are provided to allow users to enter information about their projects, experiments, specimens, and simulations. Plain text as well as other data file formats containing information about specimen configurations, instrumentation plans, research proposals, and video and images can be uploaded to many locations in the file hierarchy of the data model, using the NEEScentral upload applet. This is a bulk upload capability that allows the user to upload one or many files to a directory. Once uploaded, the researcher may add metadata (details about the file) to each individual file to enhance searchability.

4.1.2. Bulk Upload Using Excel Templates

Typical NEES earthquake engineering experiments often use dense sensor arrays on the order of 400 sensors to help characterize performance and behavior of the test specimen. Detailed metadata about each sensor, such as the exact 3-D location and orientation of the sensor relative to the specimen, is necessary in order to properly interpret and understand the data from the experiment. The NEES data model attempts to capture as much of this pertinent metadata about the sensor locations and associated data acquisition (DAQ) channels during an experiment in order to easily associate the sensors with the actual data files outputted from the DAQ system. Uploading data and metadata for each sensor and each DAQ channel individually when there are 400 sensors and DAQ channels can be extremely time consuming. NEESit has developed mechanisms for bulk upload of data and metadata to simplify the upload process for the user. One method consists of uploading many sensor locations and DAQ channels at one time using an Excel spreadsheet template provided by NEESit.

The first step involves describing essential metadata about sensors such as *sensor label, type, location,* and *orientation*. Refer to [4] for explicit instructions on how the bulk upload of sensor location metadata can be accomplished using the provided Excel spreadsheet. Once a researcher submits a filled in spreadsheet template the metadata it contains is automatically ingested into the NEEScentral database.

In addition to sensor location, it is also necessary to associate each sensor with the data acquisition channel that it is hooked up to in order to associate the sensor with the recorded time history data file. The metadata related to DAQ channels consists of the channel name/number, associated sensor, its location, and additional DAQ information such as range, resolution, and gain. Explicit instructions on how to upload DAQ channel information into NEEScentral using the excel template can also be found in [4].

4.2. Managing Users (Authentication and Authorization)

Each project can associate team members who usually belong to different organizations but are assigned to activities for the same project, thereby allowing them to collaborate as one team. Each project team member is required to register for a NEEScentral account. Information such as first name, last name, password, primary organization they belong to, role within the organization, email, address, phone and fax are collected upon registration. Each project can manage their membership and edit roles and permissions for each member. The various roles available on a project consist of *Principal Investigator, Co-PI, Grad Student, Undergrad, Collaborator, Post Doc, Research Scientist, Industry Partner, Visiting Scholar, IT Administrator, IT Programmer, Curator, Site Operations Manager, Technicians, and Other*. Each pre-defined role (e.g., Principal Investigator) is associated with a default set of permissions that control how the member is allowed to interact on the project. These default permissions are easily customized by project administrators to allow for flexible access. Permissions include View, Create, Edit, Delete, and Grant.

4.3. Advanced Search

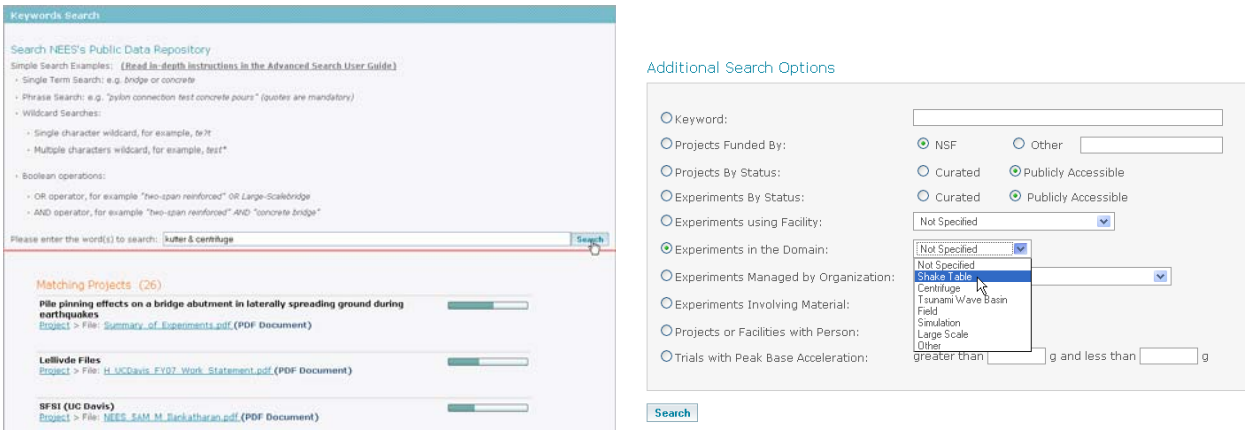
The NEEScentral Advanced Search offers users numerous options for making database searches more precise to obtain useful results. Based on Lucene technology, an open source Java product adopted by the Apache Software Foundation, the NEEScentral Advanced Search engine allows users to search the NEES database as well as supporting documents. The advances search feature provides Boolean capabilities while searching across public data sets only [8]. Figure 2a shows a screenshot of this feature.

Additionally, NEEScentral provides a canned search capability as shown in Figure 2b. Projects can be searched based on keyword, funding agency, the experimental facility or organization associated with the project and/or experiment, the earthquake engineering domain associated with the experiment (e.g., *shake table, geotech, tsunami*, etc.), the specimen material properties, a particular participant in a project/experiment, and what projects ran experiments with peak base accelerations in a certain range.

4.4. Customizable Experiment Report

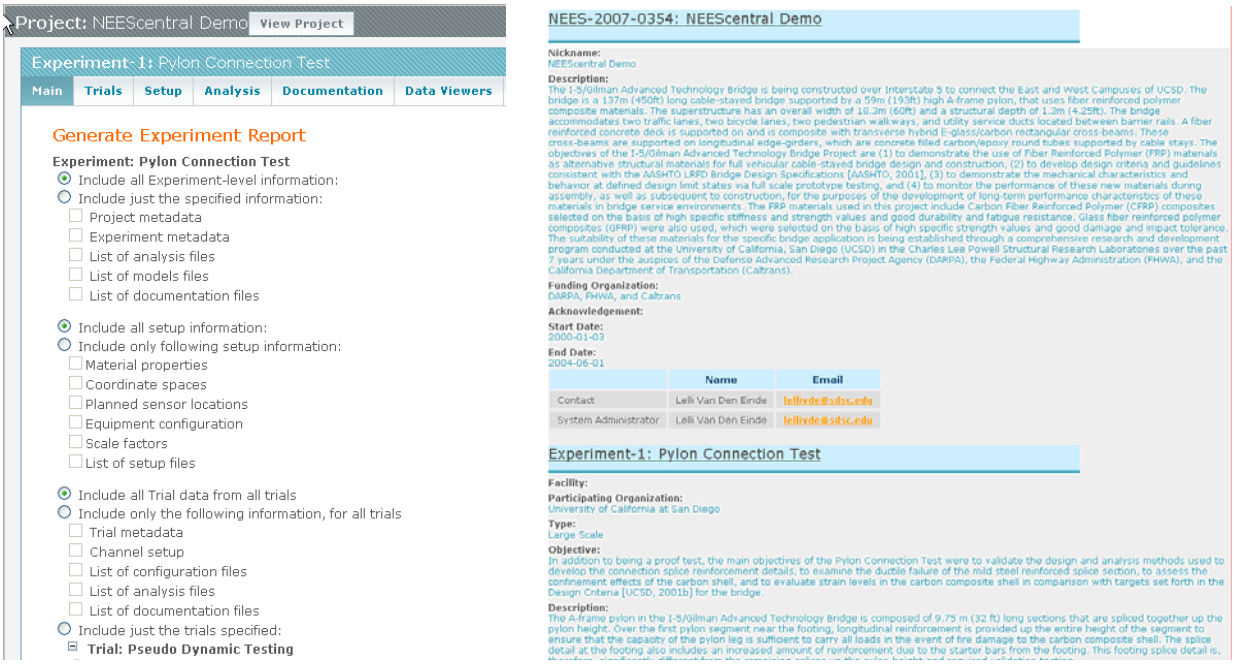
Researchers who structure their data in NEEScentral have the ability to generate customizable experiment reports. Users can select which information from their entire project hierarchy to include in the report. The content of the Experiment Report is customizable and researchers can decide what information to see which

includes experiment-level information (such as project and experiment metadata, and analysis, model and documentation files), experiment setup information (such as material properties, coordinate spaces, planned sensor locations, equipment configurations, scale factors, and setup files), trial data (such as metadata, channel setup, configuration, analysis and documentation files), and repetition data (such as unprocessed, corrected, converted, and derived data files). Additionally, researchers have the option to select the format of their output (HTML, PDF, Printer Ready Text), select how the report lists files (either as active links or static names), and select how images are displayed (either as file list entries or as expanded images). Figure 3a shows the experiment report customization page where a researcher can select criteria for generating a report. Figure 3b shows a sample HTML output.



a) Advanced Search of Public Data Sets b) Canned Search Queries

Figure 2: Advanced Search Capabilities within NEEScentral



a) Experiment Report Customization Page b) Experimental Report: HTML output

Figure 3: Customizable Experiment Report

4.5. Bulk Download

Getting data out of the repository for broader dissemination and usage is a critical requirement for the NEES data repository. Researchers can download individual data files or select multiple files for download from various locations within the project hierarchy. Additionally, researchers can download all or part of a project's

metadata and associated files into one easy-to-manage zip file. The export feature can be executed at the project level (full project export) or each experiment or even each trial can be exported separately. This allows researchers to download the data that is of interest and relevance to their own research. Figure 4

4.6. Export to UC Davis 3D Data Viewer (N3DV)

Experiment data can be exported from NEEScentral for use in N3DV, an application for visualizing experimental data developed by UC Davis [9]. Researchers are required to input specific metadata within their experiments in NEEScentral such as information about their sensor locations and channel lists that relate their sensors to the corresponding output data file. Once the required metadata and data are uploaded to NEEScentral, the application automatically generates the required input files for the 3D data viewer. The NEEScentral data repository and corresponding data model guides researchers to digitize their data and metadata that ultimately provides opportunities for the development of automated tools for visualization and data knowledge extraction, with the vision of automated pipelines for data processing into higher-order data products. Figure 5 shows an example of N3DV visualizing data exported from NEEScentral.

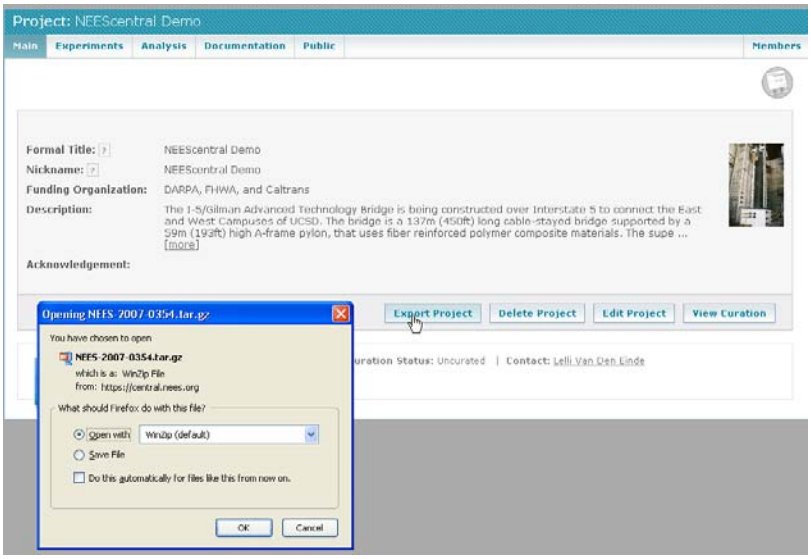


Figure 4: Export Feature

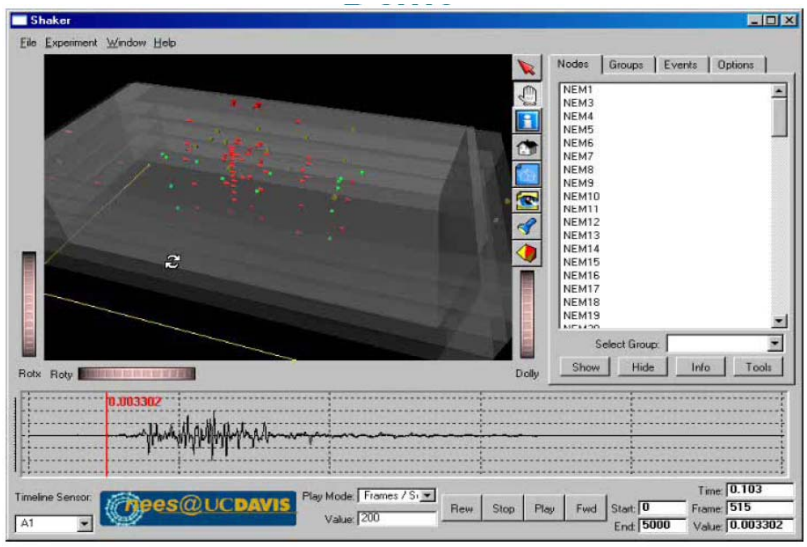


Figure 5: N3DV Visualization Tool, UC Davis

5. PROGRAMMING INTERFACE TO NEEScentral (WEB SERVICES API)

In addition to using the NEEScentral user interface, the data repository can be securely accessed via representational state transfer- (REST-) based web services. This facilitates development of applications that remotely access and control data in the repository. Communication to the NEEScentral REST Web services is initiated by the client in the form of HTTP GET, DELETE, POST, and PUT requests. GET is used to read data, DELETE is used to delete data, POST is used to create new data, and PUT is used to modify existing data. The server responds to requests by returning status codes, data, or both. Because REST transactions are stateless, each HTTP request must be accompanied by a URI that uniquely identifies the specific resource being requested.

In general, a valid NEEScentral Web services URI begins with *https://central.nees.org/REST/* as the base, followed by additional information describing the logical path to a specific resource. For example, the following URI is used to access trial three within experiment five within project 27:
https://central.nees.org/REST/Project/27/Experiment/5/Trial/3. Complete URIs can be constructed by appending successive elements and their ID numbers to the base URI, similar to the example given above. In the past, the REST based Web services for NEEScentral were implemented using PHP, and communicated with the back-end MySQL database. Currently, NEESit is working on re-implementing the Web services in Java, to communicate with the new Oracle-based back-end database. It is expected that the new Web services will be available for public use in Fall 2008.

6. RECOMMENDATIONS FOR TECHNOLOGY ENHANCEMENTS

The next generation of NEES cyberinfrastructure includes extensions to the current system in areas such as usability enhancements, integration of the NEEScentral repository into a standard portal environment, data model extensions and/or support for federated databases, and the development of a NEES digital library. NEESit is investigating easier mechanisms to get data into the system such as facilitating the upload of large data sets, easier maneuverability of data within the NEEScentral hierarchy, introducing even more flexibility in the data model, and other navigational enhancements to allow researchers to better understand and extract the content within the repository.

Furthermore, the current NEEScentral has been built as a traditional PHP-based website that provides access to the back-end central database implementation. As a traditional website, NEEScentral is fairly monolithic because all the information sources and applications were co-located. The disadvantage of such monolithic Web sites is that they make the task of introducing new applications into the framework and sharing common features across multiple projects very difficult.

With the advent of Web 2.0 technologies, the trend has been for websites to aggregate information content from diverse sources and present them in a unified way. These have come to be referred to as “Web portals”, since they provide a single point of entry or access to a set of information sources at the back-end. NEESit is moving towards a Web portals and a Service-oriented Architecture (SOA), where all the back-end components are implemented as Web services that can be accessed via a number of different interfaces. Furthermore, NEESit also plans to enable access to a federated set of databases, since different database schemas may be required to capture information from different classes of experiments. Additionally, with the use of Web portals and Web 2.0 technologies, innovative collaboration tools can be provided to the earthquake engineers such as social networking, Wikis, web forums, etc., which can be used for sharing data and early dissemination of results. More information about the proposed portal architecture can be found in [12].

7. CONCLUSIONS

NEEScentral (http://central.nees.org) is the data repository for the NEES initiative. It provides a centralized location for researchers to securely organize, store, and share data and metadata in a nonproprietary format. NEEScentral consists of an Oracle based implementation of the NEES data model, with a PHP based front-end to provide secure Web based access to the back-end database implementation. NEEScentral offers the NEES Web Services which provide secure interfaces for external access to the data repository. This enables

application developers to create applications for researchers to control the data in the repository without needing to understand the implementation details. The NEEScentral Data Repository provides data management infrastructure to shorten the gap between research and practice through long-term preservation and sharing of data. The ultimate goal is to provide access to different types of experimental and computational data to promote and facilitate collaborative and interactive processes required to address the complex nature of seismic events and their physical and societal impacts, ultimately reducing vulnerabilities from earthquakes [5].

NEES is currently working towards curation of all of the data residing in the NEEScentral Data Repository. Digital curation is the process of establishing and developing long term repositories of digital assets for current and future reference [3]. Curation should be able to show the evolution of data in the repository. It should also provide automated checklists to determine that the research data conforms to the curation standards (data completeness). NEESit would like to extend the current capabilities of the NEEScentral data archive towards representing the data as a digital library, which would serve as the public view of the data and would allow for easier search not only within the NEEScentral archive but among other federated data sources.

8. ACKNOWLEDGEMENTS

This research project is sponsored by the National Science Foundation through NEES under award number CMS-0402490. The Authors acknowledge the support of NEESinc, as well as the NEES Data User Committee consisting of Bruce Kutter, Dan Wilson, Sharon Wood, Adolfo Matamoros, Saiid Saiidi, Sherif Elfass, Gokhan Peckan, Christopher Stanton, Claude Trottier, M. Llankatharan, and A. Vosooghi. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the sponsor.

REFERENCES

[1] Whitmore, S., Van Den Einde, L., Warnock, T., Diehl, D., Hubbard, P. and Deng, W., 2006, NEESit Software Overview: IT Tools that Facilitate Earthquake Engineering Research and Education, 100th Anniversary Earthquake Conference (8th National Conference on Earthquake Engineering, 8NCEE), San Francisco, California.

[2] Buckle, I., 2005. Experimental Facilities in the George E. Brown Jr. Network for Earthquake Engineering Simulation (NEES), *Conference proceedings from the 1st US-Portugal International Workshop*, Lamego Portugal.

[3] Wikipedia, http://en.wikipedia.org/wiki/LAMP_(software_bundle), LAMP, Curation, Digital Libraries

[4] NEEScentral User Guide, http://it.nees.org/library/data/neescentral-users-guide-17.php

[5] Van Den Einde, L., et al, The NEES Data Model in Support of Earthquake Engineering Research, Companion paper for 14WCEE, Beijing, China, October 2008.

[6] Van Den Einde, L, and Kinderman, T.L., End-to-End Usage of the NEEScentral Data Repository to Facilitate Earthquake Engineering Research and Practice, Proceedings from the ASCE-SEI Structures Congress, Long Beach, CA, May 16-19, 2007.

[7] Van Den Einde, L., Kinderman, T.L, Masuda, M., Elgamal, E., NEES IT Tools to Advance Earthquake Engineering Research and Practice, Proceedings from the ASCE-SEI Structures Congress, Long Beach, CA, May 16-19, 2007.

[8] Guide to Advanced Search in NEEScentral, http://it.nees.org/library/data/guide-to-advanced-search-in-neescentral.php

[9] N3DV documentation, https://neesforge.nees.org/docman/index.php?group_id=31&selected_doc_group_id=53&language_id=1

[10] Frysinger, D., Van Den Einde, L., Warnock, T., and Agnew, G., 2006, The Curated Data Repository in Engineering Research, Paper submitted for 8NCEE Conference, San Francisco, CA.

[11] Van Den Einde, L., Masuda, M., Fowler, K., Kinderman, M., 2007, "NEESit Data Model in Support of Earthquake Engineering Research", TN-2007-05, NEESit, San Diego Supercomputer Center.

[12] Krishnan, S., et al., Towards a Collaborative Portal Environment for Earthquake Engineering, Companion paper for 14WCEE, Beijing, China, October 2008.