Curation of Earthquake Engineering Research Data

Stanislav Pejša; Network for Earthquake Engineering Simulation; Purdue University; West Lafayette, Indiana/USA
Thomas Hacker; Computer & Information Technology, Network for Earthquake Engineering Simulation; Purdue University; West Lafayette, Indiana/USA

Abstract

Earthquake engineering brings together researchers from seismology, structural, mechanical, and geotechnical engineering whose research results in saving lives and protecting property during earthquakes and tsunamis. Researchers' expectations regarding data management, data archiving, and preservation, are as different as their methodological or experimental approaches in earthquake engineering research. The diversity of the earthquake engineering community poses exciting challenges for archiving and data preservation in a domain repository, such as the Project Warehouse of the Network for Earthquake Engineering Simulations (NEES). The presented paper offers a review of the current infrastructure of the NEES data repository and further describes workflows that are pertinent to data archiving, to maintaining high quality of stored data, and to carrying out curation.

Portal for Earthquake Engineering Research

The George E. Brown, Jr. Network for Earthquake Engineering Simulation (NEES) is a research network funded by the National Science Foundation (NSF) as a member organization of the National Earthquake Hazards Reduction Program's (NEHRP), which addresses earthquake risk in the United States. Its focus is on the research community and practicing engineers who develop the innovative solutions to reduce the impact of seismic disasters.

The network consists of 14 engineering laboratories located at the top research institutions across the United States (Figure 1) that specialize in several types of experimental work: geotechnical centrifuge research, shake table tests, large-scale structural testing, tsunami wave basin experiments, and field site research.

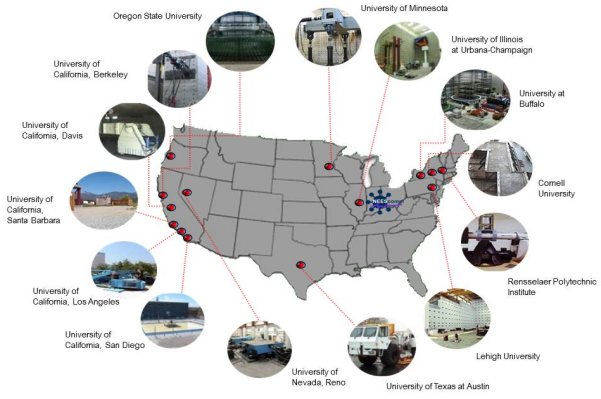

Figure 1: 14 engineering laboratories of the Network for Earthquake Engineering Simulation (NEES)

Earthquake engineering is a vibrant inter-disciplinary area that brings together researchers from seismology, structural, mechanical, and geotechnical engineering. The inherent cutting-edge and innovative character of earthquake engineering research produces an ever-changing variety of file and data formats. In this environment, confining research groups to a singular centralized metadata schema and uniform workflow for the collection of data would limit the acceptance and use of the data cyberinfrastructure by the community.

Interoperability and interdisciplinarity are desired characteristics of research conducted in the NEES network. A mandated centralized approach would not deliver the expected results in an environment where each site follows its own well-established local practices that reflect the particular strengths of each lab. Researchers, too, maintain their own internal policies developed to manage the different responsibilities of individual members of research teams that often reside in different institutions, in different time zones, and are in different stages of their research careers. To support researchers in this environment, NEEScomm chose a flexible and nimble approach to standardize where there are commonalities, and to provide flexibility for local practices in which there was no pressing need to standardize. NEEScomm's approach is based on specifying classes of materials that need to be provided, and a set of guidelines and rules that make long-term access and preservation of research data understandable to researchers and achievable given their busy research schedule.

Cyber infrastructure for data re-use

The network was established as a decentralized network of engineering laboratories each with their own local databases and only centralized points of access. The emphasis was on demonstrating and sharing the final product of the research. Gradually, the NSF and other funding agencies reoriented their focus from simple data sharing to re-use, preservation, and long-term [5, 8, 9, 10]. Following the NSF's lead, the earthquake engineering community modified the requirements for archiving and adjusted the data model underlying the NEES database and storage system.

The migration of the NEES repository to a new collaborative platform strengthened the research community's collaborative aspect. This platform, named the *NEEShub,* is a virtual research environment based on HUBzero technology [6]. The NEEShub facilitates distributed collaboration and provides access to research data and documentation necessary for understanding and interpretation of earthquake engineering research stored in

the NEES data repository named the *NEES Project Warehouse.*[1] With the change of management, the use and re-use of data became much more prominent.

The requirements for archiving [11], as well as the curation workflow were modified, so that the concerns regarding long-term access, preservation, authenticity and integrity of deposited data could be addressed. After this change, the data in the NEES data repository slowly started to take on a more consistent and predictable shape across the individual projects. Concerns over long-term preservation also drove the guidelines regarding accepted and supported format.

Archiving Data

The NEES data repository stores files collected during earthquake engineering research. These files fall into three main categories:

- sensor measurements collected from the sensors through the Data acquisition system
- data captured as still image or moving images by installed camera systems
- required documentation

Timely and safe transfer of the sensor measurements, but also of the knowledge and experience captured in the metadata and documentation are imperative for success of NEES as a network. The files can be uploaded [11] through the Web interface, called the *Project Editor*, which is a method particularly suitable for uploading documentation and providing metadata; large volumes of files such as data files or images captured by the stationary cameras during tests should be uploaded through a dedicated upload client called the *Project Explorer for NEES* (PEN). PEN uses the sFTP protocol and enables a secure and quick upload of data along with the verification of uploaded files through the use of checksums. It is also possible to use a publically available schema along with the NEES RESTful web service interface to develop an individualized client, but this method requires programmatic skills that are not expected from the earthquake engineers.

The repository's public interface with which the users interact is called the *Project Warehouse*. This is the area where research teams upload data and share their files. All files uploaded to the Project Warehouse are checked for viruses by the ClamAV software, are checksummed, and basic technical and administrative metadata are extracted and stored in the NEES Oracle database. Once the file passes the anti-virus scan, it is stored on the NEES NFS storage server. The files are then backed-up every four hours.

The real-time anti-virus check on upload slightly slows down the upload speed, and with the large number of files that can be as large as tens of gigabytes, the delay can be considerable. Therefore the format identification and validation, as well as extraction of further technical and preservation metadata is deferred to a nightly-run job, which completes the process of collecting the necessary metadata that is stored in the NEES Oracle database. The file format identification and the additional metadata are collected within 24 hours after the file is uploaded.

[1] The NEEShub and Project Warehouse are available at http://nees.org

The Project Warehouse is a *light archive*, which means that researchers can access the files they upload and there are currently no restrictions on accessing data by the nominal owners of the files. The NEEScomm documentation uses a working definition of *curated* that specifies the point at which the experimental data within a project meets NEEScomm curation criteria to preserve the content of the data and to ensure long-term access to the data. Researchers have full access to view, edit, or delete their files until their data have reached the state of *curated*. Once the experiment is curated, the members of the research team can still access the file, but they cannot edit or delete the files. Access permissions are set by the project Principal Investigator or by the IT administrator on the research team; by default access for project researchers is limited to specific experiments. Research teams are in full control of their data and own their files. They are free to upload files, delete them, and move them until the experiment is curated – at that point NEEScomm takes over control and ownership of the files. The research teams can still add documentation, analysis, or derived data, but the curated files must remain unchanged.

Metadata

In order to achieve the goals for the NEES data repository, research data must conform to best practices for their collection, delivery, curation, and annotation. Data and descriptions of objects on all levels must be based on established metadata standards. The metadata describing individual levels of research workflow, such as project, experiment, trial, and repetition are based on the feedback from the earthquake engineering (EE) community. This metadata schema was introduced by NEESinc in years 2004-2007 [12, 15, 16] and continues to be maintained and expanded by NEEScomm IT following feature requests from the EE community.

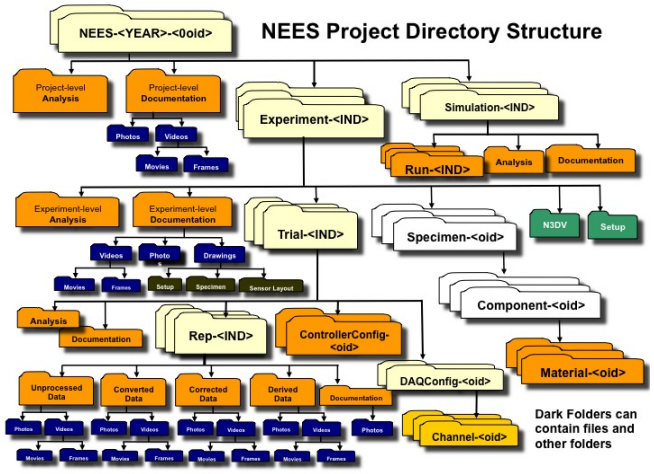

Figure 2: Simplified directory structure of the NEES file system

To ease the burden of providing metadata for all files and also eliminate the possible terminological conflicts stemming from different methodological positions of the research teams and their different research backgrounds, NEEScomm provides a directory structure where the files are stored (Figure 2). This directory structure serves as a proxy for structural metadata – these are

simply derived from the hierarchy of the directory structure of the file system. The Project Editor does not require knowledge from the researcher of the directory structure and navigates research intuitively through the individual archiving steps as researchers access various modules dedicated to archiving different aspect of their research. (Figure 3)

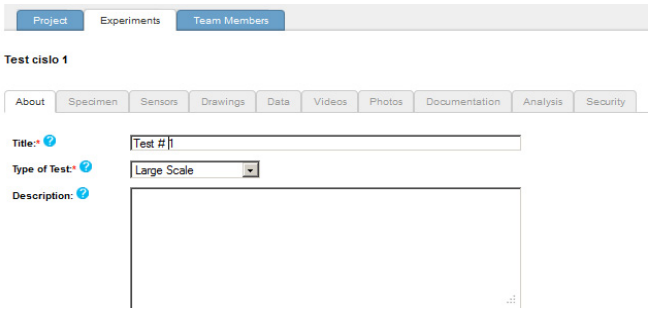

Figure 3. Tab-driven interface of the Project Editor focusing on different aspects of research workflow

The NEEScomm Requirements for Curation and Archiving of Research Data [12] identifies the metadata and documentation required of data being collected as well as supported formats for data and documentation files. Researchers can provide all files with titles and descriptions, the rest of metadata is assigned to files automatically or extracted from the files, some metadata mostly related to preservation are later provided by the curators.

The metadata schema is internal and tailored to the needs of the data administration at NEES, but it can be easily mapped to a variety of standards common in the heritage and subjects repositories. The metadata for individual files, especially those that can be contributed and shared as resources, are based on the Dublin Core Elements [4], so that interoperability with other disciplines is achieved. These objects are mainly theses, presentations, drafts of articles, etc.

The metadata for datasets are modeled on the DataCite [3] standard, which defines a minimal subset of metadata about NEES experiments, similar to bibliographic information defined for publications, so that the datasets can be published, searched, retrieved, and cited.

The preservation and technical metadata are stored in a table that is modeled on the PREMIS data vocabulary [14]. This is the most recent addition to the suite of metadata standards that was added after NEEScomm started to build up its file format registry. These metadata are based on the PRONOM vocabulary specifications [15], but needed to be slightly modified and expended to address some of the specific needs of the Network.

Quality control

The content uploaded to the repository is very diverse reflecting the experimental needs of the research teams. NEEScomm cannot dictate what content is to be uploaded as that may hamper the innovation of the research and introduce an additional burden on the research teams. Even the basic requirement to have data uploaded in an ASCII format can sometimes be difficult to satisfy across the whole repository as some domains have well established standards, yet binary, formats and requiring the data to be stored in ASCII format is not practical and would be perceived as unnecessary external administrative burden imposed on the researchers. NEEScomm tries to be pro-active and educate researchers on basic data management, but curation at NEES will always be to a certain extent reactive, as it is needs to respond to the quickly changing inter-disciplinary field, cutting edge research, innovative and novel experimental approaches.

This makes quality control and curation an essential service provided by the NEEScomm data repository. Curation is a service that assesses the fitness of the uploaded experimental output for archiving, and tries to ensure a minimum level of data quality for deposited data files and relevant documentation and mediates between the current needs of the research teams and the information needs of future researchers. The literature highlights the fact that research data management and data archiving are nowhere close to top priority for researchers who focus on conceptualizing their tests, then conducting them, analyzing the data, and disseminating their results, mainly as articles in academic and professional journals and as papers at conferences. Earthquake engineers are no exception in this respect, yet lately researchers started to refer to their datasets stored in the Project Warehouse in their presentations and to look to the Project Warehouse for available datasets for reuse.

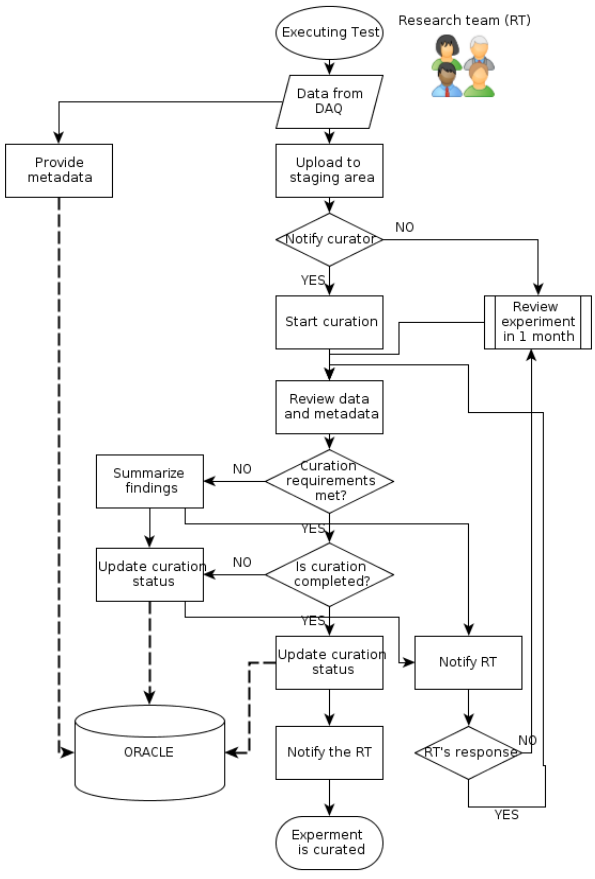

*Figure 4. **The flowchart of the curation process***

Quality assurance in the Project Warehouse is part of the curation process and focuses on two key areas: the technical quality of documentation and provided metadata, and completeness of documentation. The latter is closely entwined with timeliness of archiving. The goal is to make sure that data are archived with all necessary documentation and metadata, so that an experiment or simulation can be correctly understood and interpreted [18]. Ideally, the documentation and metadata would also allow an experiment or simulation to be repeated and its findings verified, but given the scale of earthquake engineering tests, this goal is not very realistic. In no way should curation evaluate or assess the academic merit of the given test.

The technical quality monitoring is concerned with the format of files that contain data, file integrity, preventing duplication of files, and the correct location for uploaded data.

The curators at NEEScomm try to intervene early in the experimental phase and communicate with researchers throughout the whole period of data archiving till the experimental data are admitted into the NEES Data Repository that contains curated and publically available data sets.

The curation process begins with data file upload to the Project Warehouse. A set of core metadata is collected right from the beginning, so that uploaded data can be identified and retrieved. Upon upload, researchers can provide metadata for individual files, such as title and description of the file, but most of the metadata are extracted automatically from the uploaded files and derived from the location to which the files were uploaded.

The researchers or the personnel of a laboratory typically notify the curator that their data have been uploaded. If not, the curator revisits the researcher's project in the Project Warehouse and reviews the compliance of the data with the NEEScomm curation requirements in this early stage, curators basically only note whether the data are uploaded or not. If the data are uploaded, then curators assess whether the data are in ASCII format, whether the headers are present, and whether the data are in the native Volts or whether they are converted to engineering units.

The *NEEScomm Data Sharing and Archiving Policy* [19] recognizes three major deadlines for archiving data from one test. The starting point for curation starts at the moment the research team completes their test and leaves the laboratory. Within a month, the IT personnel at the laboratory uploads the sensor measurements and some documentation to the Project Warehouse, so that the data are available to the research team who have had six months to upload corrected data and complete the necessary documentation, such as technical drawings, sensor metadata, etc. and metadata. At this point curators re-visit the archived experiment and the level of compliancy and communicate with researchers on continuous basis, so that no later than 12 months after the end of an experiment, the archived data can be curated and made public. This is also the moment when NEEScomm takes over the physical control over the files, becomes responsible for their preservation and long-term access. Upon completed curation the data are accepted to the NEES Data Repository.

While in the Project Warehouse, NEEScomm provides storage for all uploaded data that have some relevance to the earthquake engineering community. The preservation services are provided only for data that were accepted into the NEES Data Repository. The experimental data created by research teams funded by NSF through the NEES Research program are admitted to the NEES Data Repository upon successful curation; research data originating from projects funded through other agencies and within the scope of the earthquake engineering research are accepted upon approval of the curator, provided they meet minimal requirements for successful preservation and long-term access.

Formats

Making sure that the formats uploaded to the repository are suitable for preservation is also an important component of quality control in the Project Warehouse. Even if curation is re-active there is a chance to mitigate the upload of proprietary or binary data not commonly used in the earthquake engineering community during the curator-research dialog. Despite some recent advances there still seems to be a gap in recognizing and identifying a variety of research formats. The reliability of format identification is still rather questionable, so automation is still out of reach for many repositories.

Formats are a rather complex and difficult issue for NEES. Preservation best-guidelines do not easily dovetail with the cutting edge research that the earthquake engineering researchers carry out, as they often utilize new software packages or codecs that present a preservation and interoperability risk. The purpose for which data are being collected can affect the method of storage, but also how data are organized and archived.

Identification and validation are services implemented as part of the NEES preservation pipe-line that uses a stack of applications packaged as a FITS tool that is later crossed-checked manually based on additional identifying criteria. The file formats are identified as part of a nightly automated job within 24 hours after upload. This is sufficient for the curation purposes. This period typically allows enough to time to create a format profile for a given experiment and communicate any possible issues with the research team.

During the format review, the curator typically reviews files with the same file extension, and monitors the formats identified by individual identifying applications for many common formats. This is often sufficient, but for formats used for unprocessed data produced by the individual DAQ systems the investigation has to be taken several steps further. Which team uploaded the data research, which laboratory and DAQ system was involved is significant to data outcomes.

For newly uploaded data it is possible to inquire about file origin and request file uploading file in a different format. Researchers can also recommend available viewers and provide context and reasons for uploading a given class of data. The situation is trickier for files that are already in the repository. File identification offers itself as a useful tool for preservation planning and a good instrument for making educated judgments regarding the viability of maintenance and upkeep of certain classes of data. File identification is a necessary and indispensible procedure for drafting preservation plans and formating migration rules.

Files with the *.dat extension in the NEES data repository are a good example for demonstrating the usefulness as well as limitations linked to automated format identification. Files with this extension were identified as four different formats (Table 1)

- unknown binary
- Internet Explorer cache file version Ver 5.2

- Plain text
- MPEG 1/2 Audio layer 3

The least complicated category represents the files identified as plain text. Their preservation will be relatively straightforward. These files were produced by one of the laboratories. This group is the second largest.

The "Internet Explorer cache" file occurred only once in the repository. Given its location and other circumstantial evidence, we can determine that this file is not of vital importance. The *.dat files with a format identified as MPEG 1/2 Audio layer turned out to be FINDER.DAT files created when files were transferred from a Mac computer to a drive formatted for PC. These files are also not vital and can eventually be deleted and share their fate with the Thumbs.db file in the image folders.

However, the biggest class of *.dat data is the group of unknown binary. This is the category for which the PRONOM schema had to be extended, because it was also to track which research team and a laboratory is associated with certain file format and file extension.

Two subgroups were associated with DAQ systems in two laboratories – two different file formats - both of them binary and proprietary. IT is reasonable to expect that these files will outlive their usefulness by the time the whole project will be curated or certainly within a couple of years immediately after project's curation – 3 or 5 years.

Table 1: Closer analysis of *.dat files uploaded to the NEES data repository

Extension	Identified format	Manual checking
*.dat	Unknown Binary	SEED (USGS)
*.dat	Unknown Binary	UTexas DAQ
*.dat	Unknown Binary	Berkeley DAQ
*.dat	Unknown Binary	Odd files
*.dat	Internet Explorer cache file version Ver 5.2	Unidentified
*.dat	Plain text	UMinnesotta
*.dat	MPEG 1/2 Audio Layer 3	FINDER.DAT

Another large group of files originate from geo-physical research. These are binary data resembling the previous group of dat file, yet these were identified by the researcher as SEED files. SEED is a well establish standard for seismic data. These datasets can be of interest to other researcher from a seismic community and therefore should be preserved in their current format, even if it may conflict with the general NEEScomm guidelines tat require data in ASCII format.

There were several other files uploaded in this format, but the context and their location indicate that they are not vital for understanding and analyzing the data. The feedback of the community of practice regarding the policies aimed at handling these irregular files is important.

Well-formedness and validity of files as well as the fact that the extension is associated with expected format are all characteristics that are desired and are part of the quality control procedures that take place before a file is accepted to the NEES Data Repository. The fact that the JPEG file can be associated with several different formats that are all legitimate, e.g. Exchangeable Image File Format or JPEG File Interchange Format, is slightly confusing, but both formats are recognized by the digital preservation community and well documented in the PRONOM registry, but if a JPEG is identified as Extensible Markup Language (XML) or even Plain Text that is certainly worth of attention and further investigation.

Checksums created as files in the NEES workflow at several different phases of the data life-cycle, which are handled by preservation utilities as a form of verification of file integrity, can also become handy when curators try to identify duplicates that may not always be legitimate; sometimes they were uploaded by accident, sometimes they were result of confusion or misunderstanding. The duplication is also a part of quality control and part of the curator-research interchange before an experiment can be curated and admitted to the NEES Data Repository.

Conclusion – Future Developments

A key strategic criterion for further preservation solutions development for NEES is maintaining separation between the front-end platform (HUBzero) and the Project Warehouse (Oracle) where the research data are stored, so that repository transferability and portability is upheld. This approach has proven to be a successful strategy for repository transfer in the past. This separation allows for flexibility and a high-level of interoperability in case the repository needs to be moved to a different hosting environment.

NEEScomm is currently fine-tuning the maintenance procedures of the current 'light-archive' infrastructure. It expands its functionalities that extract and collect reliable provenance metadata, as well as more detailed technical, and administrative metadata. This work gradually leads towards dimming the NEEScomm data archive. File migration appears to be the primary driving force for creating the dark archive together with an effort to increase the security and safety of the stored research data and create an environment in which file integrity is secured and monitored.

Further work is planned on improving the methods for upload, improved communication between curator and the research teams. The younger researchers early in their careers are of special importance because they are typically in charge of archiving and communicate with the curator. This is also a group for which cyberinfrastructure is not something foreign and external and the curation requirements that facilitate re-use and sharing make more sense to them.

The work on format registry is ongoing and with a closer collaboration with the personnel in the individual laboratories there is a plan for building up format profiles for each laboratory. There have been several developments both in the US and in Europe [1, 2, 7] that tried to address a certain underdevelopment in the area of format identification and once the work on the local repository matures NEEScomm plans to contribute to these efforts.

Another area of intensive development is integration of the data archive with visualization and analytical tools that will assist researchers with easier discovery of data segments within the datasets that may contain patterns or characteristics worthy of further investigation.

Acknowledgement

George E. Brown, Jr. Network for Earthquake Engineering Simulation (NEES) is funded though the program

of the National Science Foundation (NSF) under Award Number CMMI-0927178. Many individuals at NEES provided opinions and corrective comments that led to this paper, but the contributions of Cheng Song especially deserve to be mentioned.

References

[1] Christoph Becker, Kresimir Duretec, Petar Petrov, Luis Faria, Miguel Ferreira, and Jose Carlos Ramalho: Preservation Watch: What to monitor and how (2012). http://www.scape-project.eu/publication/preservation-watch-what-and-how

[2] Crowd sourced Representation Information for Supporting Preservation (cRIsp) (2010), http://wiki.opf-labs.org/display/SPR/Crowd+sourced+Representation+Information+for+Supporting+Preservation+%28cRIsp%29

[3] DataCite Metadata Schema 2.2. Available at http://schema.datacite.org/meta/kernel-2.2/index.html

[4] Dublin Core Metadata Element Set, Version 1.1. Available at http://dublincore.org/documents/dces/

[5] Ixchel M. Faniel, Ann Zimmerman (2011) Beyond the Data Deluge: A Research Agenda for Large-Scale Data Sharing and Reuse . IJDC Vol. 6, No. 1, pp. 58-69 doi:10.2218/ijdc.v6i1.172

[6] T.J. Hacker, R. Eigenmann, S. Bagchi, A. Irfanoglu, S. Pujol, A. Catlin, and E. Rathje (2011). The NEEShub cyberinfrastructure for earthquake engineering. Computing in Science & Engineering 13, no. 4 (2011): 67-78.

[7] Let's Solve the File Format Problem, http://fileformats.archiveteam.org/wiki/Main_Page

[8] National Academy of Sciences. Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age (2009).

[9] National Science Board. Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century (2005).

[10] National Science Foundation. Cyberinfrastructure Vision for 21st Century Discovery March 2007.

[11] Stanislav Pejša (2012), NEEScomm Guidelines for Data Upload, https://nees.org/resources/4757.

[12] Stanislav Pejša (2012), NEEScomm Requirements for Curation and Archiving of Research Data (2012), https://nees.org/resources/4401

[13] Jun Peng, Kincho H. Law (2004). Reference NEESgrid Data Model, citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.74.5560&rep=rep1&type=pdf

[14] PREMIS Data Dictionary for Preservation Metadata version 2.2. July 2012.Available at http://www.loc.gov/standards/premis/v2/premis-2-2.pdf

[15] PRONOM Vocabulary Specification: DRAFT. version 26 October 2011 http://test.linkeddatapronom.nationalarchives.gov.uk/vocabulary/pronom-vocabulary.htm

[16] Lelli Van Den Einde, Mari Masuda, Kevin Fowler, Maritess L. Kinderman (2007). NEESit Data Model in Support of Earthquake Engineering Research. https://nees.org/data/download/NEES-2005-0067/Public/D_Data%20Model-VDE.pdf

[17] L. Van Den Einde, K. Fowler, J. Rowley, S. Krishnan, C. Baru1, A. Elgamal (2008). The NEES Data Model in Support of Earthquake Engineering Research. http://www.14wcee.org/Proceedings/files/11-0167.pdf

[18] Vermaaten, Sally, Lavoie Brian, Caplan, Priscilla: Identifying Threats to Successful Digital Preservation: the SPOT Model for Risk Assessment . D-Lib. Vol. 18, No. 9/10, September/October 2012. doi:10.1045/september2012-vermaaten on the concept of understandability

[19] Dawn Weisman, Stanislav Pejša (2011), Data Sharing and Archiving Policies, https://nees.org/resources/2811

Author Biography

Stanislav Pejša is the Data Curator at the Network for Earthquake Engineering Simulation (NEES) at Purdue University. He received his MLIS from Rutgers University (2002) and graduate certificate in Digital Information Management from University of Arizona (2010). At NEES he is primarily responsible for overseeing the quality of data uploaded to the NEES data repository and their archiving and preservation.

Dr. Thomas Hacker is an Associate Professor of Computer & Information Technology and the Co-Leader for Information Technology for the NSF NEES Project at Purdue University in West Lafayette, IN. He received his B.S. in Computer Science and B.S. in Physics from Oakland University (1989), M.S. (1993) and Ph.D. (2004) in Computer Science and Engineering from the University of Michigan, Ann Arbor. His work is focused on high performance computing and cyberinfrastructure.

