# Data Curation Profile – Structural Control

| | |
|---|---|
| **Profile Author** | J. Carlson |
| **Author's Institution** | Purdue University Libraries / Distributed Data Curation Center |
| **Contact** | jrcarlso@purdue.edu |
| **Researcher(s) Interviewed** | (Withheld) |
| **Researcher's Institution** | Purdue University / (Withheld) |
| **Date of Creation** | October 6, 2010 |
| **Date of Last Update** | |
| **Version of the Tool** | V 1.0 |
| **Version of the Content** | V 1.0 |
| **Discipline / Sub-Discipline** | Earthquake Engineering / Structural Control |
| **Sources of Information** | • An initial interview conducted on September 8, 2010.<br>• A second interview conducted on September 23, 2010.<br>• A worksheet completed by the scientist as a part of the interviews.<br>• A sample of the profiled data |
| **Notes** | |
| **URL** | http://www.datacurationprofiles.org |
| **Licensing** | Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License. |

## Section 1 - Brief summary of data curation needs

The data generated by the scientist are used to ascertain the response of a magnetorheological damper under different conditions and then to test the accuracy of a mathematical model to reproduce these responses computationally. The primary value of the data would be for other engineering researchers doing similar types of work in predictive behavior of devices and modeling. The scientist is a part of a larger research consortium of earthquake engineers, NEES Comm, which is developing an online research platform, NEES hub, which includes tools and a data repository. This data will eventually be deposited into NEES hub, but doing so will require that the data be refitted into the NEES data model and conform to NEES hub policies and standards. This refitting may include reformatting the data, providing additional metadata, and making the data compatible for use with the tools offered by NEES hub.



http://dx.doi.org/10.5703/1288284315015

## Section 2 - Overview of the research

### 2.1 - Research area focus

The scientist is an earthquake engineer studying devices that can reduce the response of buildings during earthquakes. She is engaged in a four year project with the objective of developing design methodologies that are suitable for structures that have control devices in them. She uses a new method called "real-time hybrid testing" to validate the performance of these design methodologies and their associated dampers. “Real-time hybrid testing” is a type of experiment design in which half of the system that you are experimenting on is physical and the other half is computational. During an (simulated) earthquake event the two components of this test interact with each other in real time; physical measurements are being sent to the computational model, and the computational model is sending values to the physical specimen.

The primary focus of this testing is to gain a better understanding of the devices that are being tested. The research question is testing the effectiveness of this approach as a means of measuring and ultimately mitigating earthquakes. The overall goal is to create a mathematical model for the magnetorheological damper, to be able to run the model in MATLAB and then compare the model’s results with the results gathered from the physical tests. Testing the model is done through running the model in MATLAB using the measured inputs from the experiment and then comparing the outputs generated by the model with the outputs generated from the physical experiments.

### 2.2 - Intended audiences

Researchers from mechanical and civil engineering would be interested in this data, especially those working in vibration control and non-linear dynamics and identification. Engineers in the field may also be interested in this data, though the scientist felt that researchers would be more interested than practitioners.

Researchers for the most part would be looking at models, how to reproduce behaviors computationally, or be interested making comparisons between data sets. The practitioners would probably be looking mostly at maximum force capacity; they would be looking at behavior more so than the numerical details. Practitioners would be interested in these data to get a judgment or feel for how the devices would behave or perform when they are in a structure.

### 2.3 - Funding sources

The NSF is the primary funding source for this research. As a condition of funding this research, the NSF requires that the data be shared with others outside of the project team, and that the data be preserved beyond the life of the project.

## Section 3 - Data kinds and stages

### 3.1 - Data narrative

The data consists of various inputs to drive the magnetorheological dampers in a load frame to measure and characterize their force response (output).

The test apparatus is located at the University of Illinois at Urbana-Champaign. The initial data sets are composed of measurements from the inputs to the damper as well as the resulting outputs. The inputs measured include displacement to the device, the velocity to the device, and the voltage that is applied to the device. The output primarily consists of the force response. The force is the key element of the data as it is what they are trying to reproduce computationally. Other “by-products” of the data, such as temperature values, (electric) current values, are also captured for explanatory purposes.

The data are stored on a siglab box as a vna file, which is a format used by their specific data acquisition and storage system (siglab DAQ). The data can be viewed in this format, but siglab's interface is of limited use. Therefore, the data are converted into MATLAB to make the data easier to access and analyze.

Before the data are analyzed they are cleaned. The cleaning process involves removing offsets in the data and filtering out "noise". The MATLAB data files are reviewed by a member of the project team and cleaned based upon his/her judgment. Typically any problem that arises in the data will be present in the whole test, due to an offset in a sensor or a similar problem.

In addition, a calculation is made as a part of this stage; velocity is calculated by taking a derivative of the measured displacement of the device. This calculation typically leads to a small increase (10-20%) in the file size. The number and format of the files do not change.

The data are then analyzed and plotted. The analysis is to determine is the effect of certain inputs/parameters on the data. The plots generated include comparing data inputs from within individual files such as force versus velocity, force versus displacement, or taking the results across multiple data files (as many as 10 at a time) and compiling the results into a single plot. The plots generated are distinct files themselves, separate from the MATLAB data files. The nature and the number of plots generated depend on the content of the file and what project personnel want to do with the data. Typically the compiled plots generated from across the different data files are the most useful to project personnel. The plots from single data files are not typically saved as they can be recreated fairly easily by running a script file within MATLAB. The compiled plots are generally kept as they are the ones project personnel want to refer back to over and over, and are used in presentations and publications. Plots are kept as .fig files, which contain all of the data that is in the figure as well as its properties, and can be opened and edited in MATLAB.

Data are then used to test the mathematical models that have been developed. The nature of the data and the data files are not affected by these tests. The data was gathered in the first year of the project. Data are locally stored and kept available for others in the project team. Data may be shared informally with other earthquake engineers at this point in time. All data (and some of the plots) are stored, and will eventually be ingested into a data repository, making them publicly available. The ingest process will likely require additional work to be done on the data and associated metadata to ensure conformity with the data model and standards employed by the data repository. The extent of the work required to make the data compatible with the repository has not yet been fully worked out, but may be extensive enough to warrant adding another stage to the data lifecycle.

### 3.2 – The Data Table

| Data Stage | Output | # of Files / Typical Size | Format | Other / Notes |
|---|---|---|---|---|
| **Primary Data** | | | | |
| Data Acquisition | Measurements of inputs and outputs from the test | 50 files; ~2 MB each | Siglab DAQ (.vna) | Siglab DAQ is a proprietary format, but is compatible with MATLAB |
| Conversion | MATLAB files | 50 files; ~2 MB each | MATLAB (.mat) | Conversion is done for the purposes of analyzing the data |
| Cleansing & Filtering | MATLAB files with offsets and "noise" removed | 50 files; ~2.2 MB | MATLAB (.mat) | Average file size often increases by 10-20% due to an additional calculation performed at this point |

| | | | | |
|---|---|---|---|---|
| Analysis & Plotting | Plots generated from MATLAB | Plot files: ~200 files; ~300kb<br><br>Data files: 50 files; ~2.2 MB | Plots: MATLAB (.fig)<br><br>Data Files: MATLAB (.mat) | Plots from single data files are generally not saved. Plots from compiled data are saved in MATLAB as .fig |
| Device Modeling & Data Comparison | (No new output) | (No new output) | .vna, .mat, and .fig | Data are used to test the models. All data files, and some plots are stored on a local server for local access. Stored files will eventually be ingested into NEES hub. |

**Note:** The data specifically designated by the scientist to make publicly available are indicated by the rows shaded in gray. Empty cells represent cases in which information was not collected or the scientist could not provide a response.

**3.3. - Target data for sharing**

The data from the analysis & plotting stage would be the data with the most value for others. However, the scientist is open to sharing data at any point in its lifecycle after her research results are published if asked. The scientist also stated that she will eventually make all of this data publicly available.

**3.4 - Value of the data**

The primary value of this data would be for research purposes in the engineering fields. The people interested in using this would have knowledge in an engineering field; the general public would not be able to understand this data. The scientist could imagine some educational uses, mostly to train people in earthquake engineering research techniques and practices.

This data will have value for as long as researchers continue to try and understand and predict how these devices behave. Researchers may want to identify the parameters that would be needed for their own device modeling. Device modeling is a dynamic and iterative process, models change and progress over time, but it is useful to be able to refer back to work that has been done previously.

This data could be used to make comparisons across similar devices at different locations (the data sets may not be completely the same, but will have properties that are similar enough to enable comparison). In addition, comparing data sets taken from the same device over time could be done to get a sense of how the device is changing. Over a period of time a device may change and lose some of its characteristics. Comparing data sets taken from the same device at different intervals would reveal changes in the device.

**3.5 - Contextual narrative**

This project is affiliated with NEES Comm, a shared network of 14 experimental facilities supported through an NSF grant. NEES Comm operates the NEES hub, a virtual community that includes a repository for data generated by NEES Comm projects. Data within this repository are meant to be organized in such a way as to make them compatible with each other. Data organization, management, curation and preservation are governed by the NEES data model. The NEES data model has been made available to project personnel and is currently in the process of being implemented.

The scientist's data is currently in the process of being transferred from a server at UIUC into the NEES data repository and fitted to conform to the NEES data model.

The scientist is one of the developers of the NEES data model and serves on the NEES data working group. She has previously submitted data from other projects that she has been involved with into the NEES hub repository.

The scientist stated that the "culture" of Earthquake Engineering community is beginning to move towards a greater understanding and acceptance of sharing of research data. This transformation is partially due to the rules imposed by the NSF and partially due to the fact that the research community is recognizing that if these data sets were made available more could be done with them. Naturally, there are some researchers who will never share their data unless they are forced to do so; however there are also a number of researchers, particularly those who are in their first 15-20 years in a faculty position who are excited about sharing the data. Typically, researchers are more willing to contribute their data because they feel like others are going to do this as well. One of the goals of NEES hub is to serve as a mechanism to facilitate data sharing amongst the Earthquake Engineering community.

## Section 4 - Intellectual property context and information

**4.1 - Data owner(s)**

Ownership of the data has not been clearly established. the scientist feels that she owns the data as the PI for the project, in conjunction with the researchers at the University of Illinois at Urbana-Champaign as they are the ones generating the data at their facility. Ultimately, she would attribute ownership of the data to the project team as a whole, which includes personnel from five different universities.

**4.2 - Stakeholders**

The NSF, as the primary source of funding for the research generating the data, is a stakeholder for this data. The NSF as a condition of funding has required the scientist to share her data with others and make arrangements to preserve the data beyond the life of the project.

**4.3 - Terms of use** (conditions for access and (re)use)

In sharing her data with others, the only condition that the scientist would ask is that anyone who made use of her data to agree to acknowledge the source of the data set in any resulting research paper.

This data is not subject to any privacy or confidentiality restrictions or concerns.

**4.4 - Attribution**

The scientist would expect attribution from others who made use of her data set; this is a high priority for her. The ability to cite this data set in her publications is also a high priority for her.

## Section 5 - Organization and description of data (incl. metadata)

**5.1 - Overview of data organization and description (metadata)**

Information about the conditions under which the data were gathered are recorded manually by project personnel at the time of the test and listed in a text file (a "read me" file); separate from the data itself. Other elements included in the file are some of the temporal aspects of the experiments, which day the experiment was conducted and what kind of tests were performed on a particular day for example, and notes on any anomalies that may have occurred. No tools to automate the collection of metadata are currently in place.

Data are often gathered through repeating trials on the same device with minor variations of the calibrations. The "read me" file contains these calibration constants across data files. The "read file is typically stored in the same folder with the relevant data sets so that people know which calibrations are associated with which data sets. In addition, data files are named according to a particular naming convention that enables the calibration information listed within the "read me" file to be associated with a particular data file (or files).

In addition, when the data and metadata are submitted to the NEES repository, project personnel must also provide the factors to do any conversions required for the metadata to conform to the NEES data model. For example, the units of analysis in the data set may be needed to be converted once the data are uploaded in order to conform to the NEES data model.

The scientist feels that this data set is sufficiently described and organized that another researcher in her field could understand and make proper use of them.

**5.2 - Formal standards used**

The scientist indicated that the ability to apply standardized metadata from her field to this data set is a low priority. Two reasons were given for this response. First, she feels that her data has a relatively small number of attributes and is fairly straight forward in comparison to some others in her field. She indicated that the time and effort it would take to apply a formal standard to her data set would not likely be justified by the potential benefits.

Second, she stated that currently "standardized metadata" in her field is a fuzzy concept. There are no formal standards in Earthquake Engineering that would be readily applicable for this data set. The research being conducted in the Earthquake Engineering field is diverse enough that a single standard would likely be insufficient; three or four standards may need to be developed.

**5.3 - Locally developed standards**

For this particular data set, the calibration of the device and other information in the "read me" field are straight forward enough that they can be easily understood by project personnel. The names of the data files follow a naming convention so that the files can be associated with information in the "read me" file.

Within the NEES hub, the data model will define how the data should be described and formatted. The NEES hub requirements in these areas are complex as they must accommodate the needs and practices of individual researchers while developing a larger, interoperable collection of data. The details of the data model on formatting and description are still being worked out amongst NEES hub administrators and affiliated researchers. NEES hub personnel are considering implementing data format converters as a part of the submission workflow.

Depending on the decisions made and how the data model develops, the scientist's data may need to be converted from MATLAB into a different format as a part of its processing into the NEES hub. Use of this data with some of the tools in NEES hub may also require changing the format of the data.

**5.4 - Crosswalks**

Information in the "read me" file may need to be refitted or recalculated to come into alignment with the NEES hub data model. It is unclear to what extent the scientist's data requires this kind of realignment.

**5.5 - Documentation of data organization/description**

These data are described primarily by "read me" files which are stored in the same folder as the data sets it describes.

Once this data set is submitted and fully integrated into the NEES hub repository it will conform to the NEES data model.

## Section 6 - Ingest / Transfer

Currently, the scientist and the project team are in the process of transferring this data from a server at UIUC to the NEES hub, a centralized portal for researchers engaged in earthquake

engineering. The services provided by NEES hub include a data repository and tools for analyzing or visualizing data.

Once uploaded into the NEES hub, project data are transferred to one of four folders by the person uploading the data based upon where the data are in its lifecycle. There are folders for unprocessed data, converted data, corrected data and derived data. The scientist's data has been uploaded to the unprocessed data and the converted data folders. For this data set, they probably will not have "corrected" and "derived" data, as defined by the NEES data model.

In addition to the data, project personnel provide the metadata, including any measurement conversions or other information necessary to populate the data model, as a part of the ingest process. The scientist currently provides this information for her data in a separate text file that contains the needed values and information. This text file was uploaded to the NEES hub as a part of the data transfer process.

An important element of the NEES hub repository identified by the scientist is that uploading (and downloading) data should be straight forward and a relatively easy process. (1-54:20) There have been some questions from project personnel about how to follow the NEES hub data model, specifically on how to classify their data and how to determine which folder they should deposit their data. These questions have led to some difficulties in completing the curation process for some of the data sets submitted to the NEES hub.

Data are generally reviewed and ingested into NEES hub through an informal process; however the scientist is seeking to standardize this process in the near future. The formal process will likely consist of the researcher who conducted the test to download his/her data set from the hub and spot check it. The hub's data curator will do some high level checking of the data after ingest as well.

The ability to batch load data files into the NEES hub repository is a high priority for the scientist.

## Section 7 – Sharing & Access

### 7.1 - Willingness / Motivations to share

According to the terms and conditions imposed by the NSF on her award, the scientist will need to release this data publicly after 12 months. Generally she would be willing to share data anyway, even without the NSF requirements, if she were asked to do so. The scientist indicated that she would be willing to share the data with her immediate collaborators right after the data had been generated. She would be willing to share her data with other earthquake engineers after the data has been processed for analysis. She would be willing to share the data with researchers outside of her field immediately before publication, and would be willing to share her data publicly once her research findings were published.

People within her research team are currently analyzing the data and using it to develop their models. Data has not yet been shared outside of project personnel as it was generated only last year. The scientist and others have not yet presented or published all of their research findings, although several papers will be submitted in the near future.

The scientist has not yet made this data available, nor has she been contacted by anyone seeking access to her data. She attributes this to the public availability of a similar data set generated a few years earlier by one of the projects co-principal investigators, reducing the immediate need for her data. The scientist imagines that other researchers may want to compare her data with this existing data set once her data are made available.

In previous instances where she has shared her data it has been more of an informal process of emailing the data files themselves and perhaps some descriptive information about the data. The

scientist is looking to develop a more formalized process to share her data publicly using the NEES HUB research portal (http://nees.org/). She has already submitted other data sets that she has generated into NEES hub for the purpose of sharing them publicly with others.

**7.2 - Embargo**

The scientist indicated that she would require an embargo of about 1 year for this data set. This is the standard given by the NSF, which was the main factor behind her response to this question. It was somewhat difficult for her to give a precise response as the time it takes to analyze and develop a model based on the data can vary from project to project.

**7.3 - Access control**

Once the data has been released from its embargo it can be made freely available. No additional access controls beyond the 1 year embargo period are needed for this data set.

**7.4 Secondary (Mirror) site**

The ability to access the data set at a secondary site if the repository is off-line is a medium priority for the scientist. She is assuming that the NEES hub (the eventual repository for her data) will only be offline for short periods of time (hours rather than days) for maintenance and upgrades.

## Section 8 - Discovery

Ideally, the scientist would like to have her data's metadata captured by the NEES hub in electronic form, rather than as a flat file, and used to augment the search and browsing capabilities of the hub. The type of testing done to generate the data (sine wave testing, etc.) was specifically mentioned as a attribute that others would want to use to search or browse her data and other data sets like it.

The scientist indicated that this data set would primarily be used by other Earthquake Engineers, therefore the ability for researchers in her field to find her dataset easily is a high priority for her. Other engineers using a magnetorheological damper or similar device may find this data to be useful so their ability to discover it would be a medium priority. Discovery of this data set by other audiences is a low priority.

The primary means of discovery for this data would be through the NEES hub, however the scientist also indicated that the discovery of this data set through internet search engines was a high priority for her.

## Section 9 - Tools

The scientist and her project team make heavy use of MATLAB in converting, analyzing and plotting this data set. They would like to see the MATLAB tools they use on a regular basis incorporated into the NEES hub environment, particularly the plotting, simulation and analysis tools. The scientist recognizes that incorporating MATLAB tools into the hub presents a challenge particularly as her team uses such a wide variety of the MATLAB tools in their work. NEES hub does offer a variety of useful tools in some areas, such as plotting, however adjusting from the familiar interface of MATLAB to the new interface provided in NEES hub takes some getting used to.

NEES hub has developed a data model to govern the submission, processing, handling and treatment of the data within the repository. However, this data model is fairly complex and researchers have not yet received much training on the data model. This has led to some confusion on the part of researchers over the process of submitting their data to the NEES hub.

One suggestion to address this problem was a series of short video clips to train personnel on the purpose and use of the data model, the process of submitting data and the functionality of the NEES hub repository.

Due to the diversity of data being targeted for collection by the NEES hub there may be a need for the development of more than one data model to accommodate them all effectively and efficiently.

The ability to connect the data to visualization and analytical tools is a high priority for The scientist. The NEES hub offers such tools and continues to develop new tools for making use of earthquake engineering data. Using some of the tools may require that her data be converted into a different format, specifically from MATLAB into ASCII.

The scientist could imagine that there might be some cases in which researchers would want to comment on her data set, making this a medium priority.

## Section 10 – Linking / Interoperability

The scientist indicated that connecting her data to any resulting publications was a medium priority for her. Currently, journals in her field do not generally accept data or other supplemental information for publication.

She also indicated that other researchers may want to use her data set in conjunction with other similar data sets, although she did not believe that this would be a primary use of her data. For her personally, merging her data with other data sets is a low priority.

Her data set is currently being added into the NEES hub data repository. The data model developed by the NEES hub seeks to ensure that the data added to the repository are formatted and described in a manner that enables the discovery of all data in the repository and that the data can be plugged into the NEES hub tools.

The NEES hub's ability to support the use of web services is a low priority for the scientist.

## Section 11 - Measuring Impact

### 11.1 - Usage statistics & other identified metrics

The NEES hub is capable of providing download statistics, but the scientist was unaware of how much detail is captured in the statistics or if this functionality had been fully implemented yet.

### 11.2 - Gathering information about users

The scientist would like to know who has published research that made use of her data in some way. Having information on publications generated from her data would be more interesting and useful to her than knowing the number of times the data was downloaded.

## Section 12 – Data Management

Previously the data was stored on a network drive over at NCSA at UIUC. The data was made accessible to all project team members (not just those at UIUC) using a login and password. The data is being transferred to the NEES hub in preparation for their public release; however the organization and description of the data are not yet complete.

**12.1 - Security / Back-ups**

Data stored on the network server at UIUC are password protected

The data are currently backed up to local hard drives. Locally, the frequency of back-ups is determined by event rather than time. When new files are added or significant changes are made the data are backed up. NEES hub is backed up on a nightly basis. Copies of the backed up data are stored in 3 or 4 different geographic locations.

The scientist is aware that security measures to protect her data and other content on the NEES hub, but she is not sure as to what the specific measures are. Security is not much of a concern for her and she trusts the IT people running NEES hub to provide a secure environment.

**12.2 - Secondary storage sites**

Back-up copies of the data are kept in different geographic locations. Locally, one copy of the data is kept at Purdue, one at UIUC, and one copy on the NEES hub. Back-ups from the NEES hub are stored in 3 or 4 different geographic locations.

**12.3 - Version control**

Although new files may be added to the data set, once the data are generated they do not change. Therefore, the scientist rated the ability to enable version control for her data as a low priority.

## Section 13 - Preservation

The scientist has saved most, if not all of the data that she has generated since graduate school. However, she has not taken steps to migrate this data out of older or obsolete mediums or formats, and so she is not sure if she could still access some of her data sets. She is looking to avoid this situation with this data.

Even though the data at the analyzed stage of its lifecycle is likely to have the most immediate value for other researchers, the scientist believes that the raw data are the data that should be preserved from her data set. Different researchers may have different perspectives on how to work with her data, including how this data should be corrected or filtered. It is the raw data that will always serve as the baseline. The processing and analysis stages can be reproduced from the raw data if needed.

Secondary storage sites for the data as a part of the preservation activities is of high importance to the scientist.

**13.1 - Duration of preservation**

The scientist believes that her data should be preserved for 10 years or more but less than 20 years. The scientist believes that the mathematical model they are working on will be perfected within this timeframe. When the model is considered complete the data will lose much if not all of its value. In addition, the device used in generating the data may become obsolete within this time period and replaced, again diminishing the value of the data.

**13.2 - Data provenance**

Documentation of any and all changes made to the data over time is a high priority for the scientist.

**13.3 - Data audits**

The ability to audit this data to ensure its integrity over time is a medium priority for the scientist. She does not want the responsibility of auditing the data herself, but she would like to be sure that audits are being done by someone.

**13.4 - Format migration**

The raw data that the scientist identified as needing to be preserved are currently in a proprietary format, .vna. The scientist stated that the data should be migrated out of .vna format for preservation purposes and into a more common format such as MATLAB or ASCII.

The scientist indicated that the ability to migrate this data set to new formats over time is a medium priority.

## Section 14 – Personnel

**14.1 - Primary data contact (data author or designate)**
(Withheld)

**14.2 - Data steward (ex. library / archive personnel)**
(Withheld)

**14.3 - Campus IT contact**
Unknown

**14.4 - Other contacts**
(Withheld)