NHERI Data Curation Guidelines

Introduction

This document provides data curation guidance and best practices for researchers who will use the NHERI DesignSafe cyberinfrastructure (CI) to share and publish natural hazards engineering data. Specific guidelines are provided for researchers using the NHERI Experimental Facilities (EF), for researchers using the NHERI RAPID facility, and for researchers performing simulation-based research.

Data curation is made up of all the activities undertaken to generate organized and documented data that is easy to re-use. Using data management tools in DesignSafe, researchers are empowered to progressively curate their own data as their research progresses. On demand assistance from a curator is available to provide training and to guide users through their data curation and publication needs. When curation is complete, researchers can publish the dataset with a permanent digital object identifier (DOI) that allows the data to be easily located on the web and cited. Features are in place to ensure the authenticity, integrity, security and persistence of the datasets for open access. DesignSafe is committed to the continuity of data preservation beyond the conclusion of the DesignSafe project.

To cite the use of DesignSafe in your research, please reference the following paper:

> Rathje, E., Dawson, C. Padgett, J.E., Pinelli, J.-P., Stanzione, D., Adair, A., Arduino, P., Brandenberg, S.J., Cockerill, T., Dey, C., Esteva, M., Haan, Jr., F.L., Hanlon, M., Kareem, A., Lowes, L., Mock, S., and Mosqueda, G. 2017. "DesignSafe: A New Cyberinfrastructure for Natural Hazards Engineering," ASCE *Natural Hazards Review*, doi: 10.1061/(ASCE)NH.1527-6996.0000246

Data Sharing and Publishing

DesignSafe provides an end-to-end data management, analysis and publication platform for both experimental and simulation-based research. Within the DesignSafe Data Depot, researchers have access to a private "***My Data***" space, a collaborative "***My Projects***" space, and a "***Published***" space for published datasets.

All research data collected as part of a research project, as well as processing scripts, data analysis products, and simulation models/results generated, can be deposited in the Data Depot from the inception of the project. These data are kept private within a ***Project*** until published by the research team. Using a ***Project*** to share data with your team members during the course of a project facilitates the progressive curation of data and eventual publishing.

Each research team is responsible for curating its data using the data management tools provided by DesignSafe. These tools help researchers organize, categorize, and describe their data within the DesignSafe Data Depot. Assistance from a curator is available to provide training and to guide users through the data curation process. After data is curated and ready to be published, it will be vetted against the research community's minimum metadata requirements (https://www.designsafe-ci.org/rw/support/data-publication) before moving on to receive a DOI for persistent identification and ease of data sharing and citation. Researchers using published

data from the DesignSafe Data Depot must cite it using the DOI which includes the DataCite schema for accurate citation (http://schema.datacite.org/).

Responsibilities and Timelines

Researchers working at a NHERI EF will receive their bulk data files via the Data Depot. NHERI EF staff will deposit the data files into an existing ***Project*** created for the research project. For all other types of research (e.g., simulation, experimental work performed at a non-NHERI lab), it will be the responsibility of the research team to upload their data to the Data Depot. As noted previously, the research team is responsible for data curation and publishing. Although no firm timeline requirements are specified for data publishing, researchers are expected to publish in a timely manner. Recommended timelines for publishing different types of research data (i.e., Experimental, Simulation, and Reconnaissance) are listed in Table 1.

Table 1. Recommended Publishing Timeline for Different Data Types

Project/Data Type	Recommended Publishing Timeline
Experimental	12 months from completion of experiment
Simulation	12 months from completion of simulations
Reconnaissance: Immediate Post-Disaster	3 months from returning from the field
Reconnaissance: Follow-up Research	6 months from returning from the field

Licensing

Within DesignSafe, you will choose a license for your material. Because the DesignSafe Data Depot is an open repository, the following licenses will be offered:

- For datasets: ODC-PDDL and ODC-BY
- For copyrightable materials (for example, documents, workflows, designs, etc.): CC0 and CC-BY
- For code: any open, non-commercial license (for example, GPL)

You should select appropriate licenses for your data after identifying which license best fits your needs and institutional standards. Note that datasets are not copyrightable materials.

Data Archiving and Preservation

Depositing your data and associated research project materials in the DesignSafe Data Depot will meet NSF requirements for data management. DesignSafe will persistently maintain all uploaded data on storage resources at the Texas Advanced Computing Center, and these resources are redundant and geographically replicated. DesignSafe operates a dedicated Fedora repository platform to ensure the authenticity, integrity, security and persistence of published datasets for open access.