Best Practices for Publishing Reproducible Simulation Datasets on DesignSafe

Draft 3.0 (August 2023)

DesignSafe Simulation Publication Working Group: Clint Dawson, Maria Esteva, Jamie Padgett

Introduction

This document provides best practices for publishing simulation models and datasets in DesignSafe.[[1]](#footnote-0) The topics addressed were extracted from literature reflecting the modeling community needs and recommendations [1, 2], and informed by the experience of the working group members and the larger DesignSafe expert cohort while conducting and curating simulations in the context of natural hazards research [3].

Simulations are complex processes, each with unique characteristics, which presents curation and publication challenges. Acknowledging that it is difficult to develop a “one size fits all” set of best practices, we chose **reproducibility** as the key concept to generalize these recommendations. According to Peng's definition of reproducibility [4], an independent researcher needs a complete dataset, the computer code used to generate/read/analyze it, and documentation for both in order to verify the results. The challenge with large simulations is to determine and normalize what constitutes a complete dataset, which and how many files need to be published, which software and ancillary codes are indispensable, and the level of documentation needed for all. Overall, these best practices emphasize a complete simulation publication including precise reference and description of the model, and complete publication of inputs and when possible all outputs.

In crafting these best practices we considered the characteristics of simulations conducted within the natural hazards space, and the functionalities available in the simulation curation and publication pipeline in DesignSafe. This is a working document that will evolve as we observe how best practices are implemented, and gain experience from new publications and community feedback.

Best Practices

# DesignSafe Simulation Data Model

For purposes of normalizing simulation data publications, DesignSafe has implemented a simulation data model which includes the following categories to organize the data:

* Project: A project management or administrative umbrella under which one or more simulation datasets can be grouped. For example, different simulation datasets generated under one research project performed and funded through the same grant.
* Simulation: A stand-alone dataset published under a unique citation and DOI. One or more can be published under a project and each can be published progressively and versioned.
* Model: Files and/or information describing the design, geometry, and/or code of a simulation. This includes referencing a community model and version when applicable.
* Input: Files or references to the files containing the parameters of the simulation.
* Output: Files containing the results of a simulation.
* Report: Detailed written accounts made to convey information about the simulation. It could be one or a group of files including a readme file and a data dictionary, depending on the complexity and components of the simulation.
* Analysis: Tables, graphs, visualizations, Jupyter Notebooks, or other representations of the results.
* Software: (to be released in 2024) Code and documentation containing the research software used in the development or analysis of the published dataset. This includes a unique citation.

Each category includes required or recommended metadata elements that help characterize them. We require simulation publications to include metadata and files in the Model, Input, Output, and Report categories. The Analysis category is optional. This model constitutes the baseline for researchers to publish datasets for reproducible research.

# Describing Simulations: Titles, Descriptions, and Keywords

*Titles*, *descriptions*/*abstracts*, and *keywords* are written for purposes of communicating key aspects of the dataset publication, for making your datasets understandable and for advertising the datasets. These descriptive elements are presented in the landing pages of DesignSafe's Data Depot and are searchable on the web and thus fundamental to the discoverability of your dataset on the Internet. General recommendations to write them are:

* Use language that can reach experts as well as layperson audiences.
* Use words that you consider will help others find your dataset online.
* Repeat these words across the title, description and keywords to increase online discoverability.
* Avoid or define all acronyms used.
* Descriptions/abstracts should be between 200 - 250 words.
* Keep descriptions concise and engaging, further details about the dataset will be introduced in a Data Report.

DesignSafe simulation publications have two levels of description: project and dataset. The reason for this is that one project may have more than one dataset publication so the project is an umbrella to contain various dataset publications. These two descriptive instances should be complementary, the project description giving a broad description of the research and the dataset description on the generation and characteristics of the data. Data should be described as a standalone research output, so it can be understood independently from related research products such as a published paper [5].

## Simulation Project

The project is intended to communicate the broader goal of the research, and who the intended audience is.

Project Title

* The project title can be similar to that of the grant project that supports the research and it should be distinct, encompassing the title(s) of the datasets.

Project Description

* Begin the description/abstract with a general statement that provides context to the study (e.g., the system under investigation).
* Describe the type(s) of hazard being studied (wind, earthquake, wildfire, heat, etc.).
* Use direct language, **do not copy the abstract of the proposal**, as that describes what the project will do and not what has been done.
* The data is the focus of the publication; describe the function of the data in the project.
* Indicate if there are more than one datasets in the project, and if more than one institution is involved.
* Indicate who will benefit from the data.

## Simulation Dataset

The following are recommendations to write the simulation dataset description.

* To help focus on the dataset you may begin the description with "This dataset…"
* You may repeat some concepts from the project description as needed.
* Indicate the model used to conduct the study and cite it in Related Work.
* Describe the structure of the simulation (e.g. if it has more than one Model, multiple runs with different configurations, if it includes analysis files, etc.)
* Indicate the type of inputs used (you can cite the input sources precisely in Referenced Data).
* What was the outcome?
* How can the data be reused (reproducibility, generate new studies, validation, ML, etc.)?
* Further details about file formats, file naming conventions, results
* If you want to direct users to papers written about the dataset, include those in Related Work using the tag “Cited by”.

### Keywords

* As descriptive elements, keywords are used by other researchers to find your datasets online.
* When applying keywords researchers have to think how others would search for their particular dataset.
* As applicable, use keywords to indicate: type of hazard, system being simulated, location, community model used, problem addressed, purpose of the simulation, and results.
* Repeating words used in the description and titles increases the chances for the dataset to be discovered online.

# Simulation Categories: A Complete Representation of the Dataset

Following are recommendations regarding the descriptions and the types and number of files to include in each category of the simulation dataset. Categories convey the structure of a simulation by showing the components and the way in which they were configured to run the simulation. They also ensure completeness so that all the data, code, and documentation required to reproduce it is available. A simulation may have different configurations and DesignSafe’s data model allows flexibility in representation. Figure 1 below shows how a simulation was organized using the simulation model categories in DesignSafe.

* When creating a category title, be descriptive and make it unique from other categories. Use sequential ordering if necessary.
* Summarize the purpose of the category and its files. What is it about? What are its features?
* Before uploading the data to the Data Depot, it is helpful to organize it following the categories defined in the data model.
* For quick access to the files by a viewer, avoid excessive nesting of files deep in folder structures. Instead, use the categories as the root container for files.
* If data needs to be grouped further, use folders with clear file naming conventions (see recommendations for file naming conventions)
* After grouping files in categories, select pre-defined DesignSafe file tags or create custom tags to describe the files. Tags should reflect each file's contents, context, and purpose.
* Refer to the DesignSafe curation and publication User Guide for more details on how to publish simulations. The guide is located here:   
  <https://www.designsafe-ci.org/user-guide/curating/>

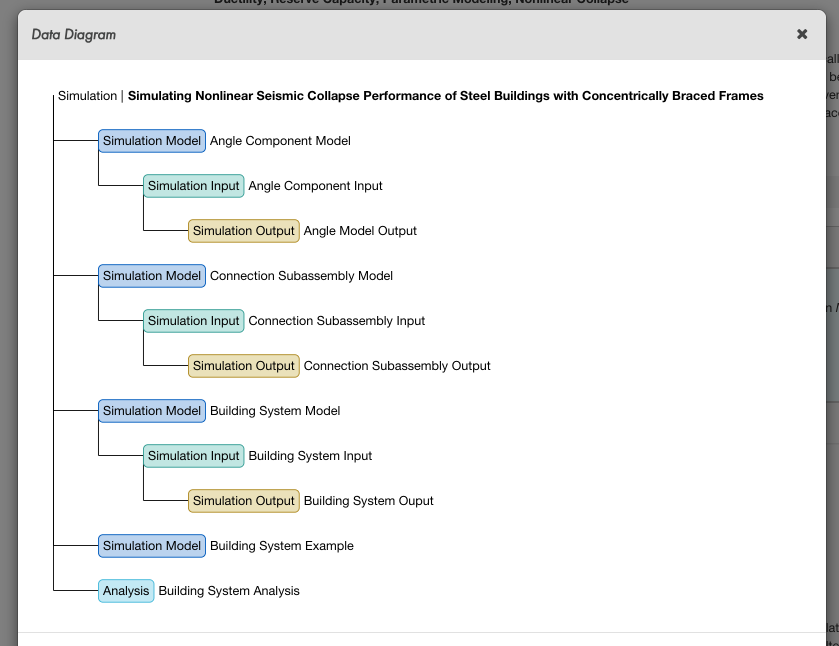


Figure 1. DesignSafe categories to curate and organize simulation datasets in a tree view for the following simulation: Bradley, C. (2021) "Simulating Nonlinear Seismic Collapse Performance of Steel Buildings with Concentrically Braced Frames", in *Modeling and Analysis of Steel Frame Building Systems*. DesignSafe-CI. <https://doi.org/10.17603/ds2-hwdv-c789>

Other more complex simulation publications:

Asher, T., (2019) "Hurricane Matthew Storm Surge and Wave Simulations", in *Hurricane Matthew Storm Surge and Wave Simulations with Data Assimilation*. DesignSafe-CI. <https://doi.org/10.17603/ds2-ne6j-s897>.

### Model

* The model category description is required to express the design, geometry and or software of the simulation (, as well as inclusion of files .?)
* Describe the software used in the simulation (community model, research software, proprietary software, etc.). See the Software section below with recommendations of how to document it.
* When applicable, include information about spatial and temporal scales and resolution. This information should also be included in the Report (See Report Section).
* When applicable, add information about the computational environment used to run the model.

### Software

When using published community software:

* Use the field Referenced Data and Software to note software name, version, and the DOI or URL where the current version of the software is located.
* Include all ancillary scripts and files required to run the community code and reproduce the results – including those published in a paper or report.
* Include a readme file to explain how to run the community software.
* Use available or custom tags to indicate the purpose of the files.

When researchers created the software:

* If the software is published elsewhere with a DOI, reference it in the Referenced Data field.
* If the software has not been published elsewhere, authors can publish it using the Software category in DesignSafe (to be released in 2024). To be published, the software should comply with our definition of research software.
* Include all ancillary scripts and files required to run the software and reproduce the results of the simulation dataset you are publishing.
* Include a readme file to explain how to run the research software.
* Use available or custom tags to indicate the purpose of the files.

### Input

Describing, publishing, and referencing the input files used in a simulation are essential for reproducibility. Researchers should comply with the following best practices:

* DesignSafe requires making the provenance of the input files clear so that users can access these files to reproduce the simulation. This can be achieved by referencing and publishing the input files.
* The public source of the input files should always be recorded in the Referenced Data field including a DOI or URL. This provides credit to the agency/author providing access to the input files.
* Republishing the input files is always convenient for other users so they can reproduce your research promptly (it reduces time for them to find the inputs elsewhere).
* When referencing input files it is important to point to the precise files. Considering that DOIs or URLs may point to landing pages but not to the precise files, researchers should describe how the input files can be retrieved. For example, researchers may need to include the query by which they were retrieved from a database, or the directory and file path locations, or file naming conventions.
* In cases in which the source of the files is known but the precise location is not available, (or it is practically impossible to point directly to them), the alternative to facilitate reproducibility is to republish the files in DesignSafe.
* Proprietary input files that cannot be republished due to original usage agreements or restrictive licenses should be cited in the Referenced Data field including a DOI or URL. If such pointers do not exist, the provenance of the proprietary files should be described in the Simulation description with as much precision as possible in case someone needs to access them by contacting their creator.
* Researchers should check the distribution licenses of the input data to make sure that they can be modified and republished.
* When public datasets do not have explicit licenses, there is a tacit understanding that these files are in the public domain.
* If input files have been pre-processed for usage in the published simulation, the processed files should be published including a description of the pre-processing method. Researchers may determine if they should make both raw and processed input files for reuse. Distinction between raw and processed should be clearly noted.
* Input files may have obscure naming conventions that are functional to the model/software used in the simulation. Use the pre-defined file tags to describe the content and purpose of the input files.

### Output

Simulation outputs can be significant in size, which brings up the question of whether to publish all of them. In Open Science Expectations for Simulation-Based Research [1], simulations are classified as for knowledge production or for data production. The former are intended to conduct experiments and publish results; the latter for data reuse to generate more knowledge. In relation to publishing data, the general idea is that for simulations created for purposes of data production - which can be reused as benchmark and testbed datasets - outputs should always be published. On the other hand, the outputs of knowledge production simulations may not be published if the elements required to re-run the simulation (inputs, model used, scripts, etc.) are available and documented.

After careful considerations, DesignSafe recommends and facilitates publishing output files belonging to both knowledge and data production simulations, as data outputs are always required to conduct reproducible research. Arguments in favor of publishing outputs include the cost and difficulties involved in re-computing large-scale simulations. At the moment, DesignSafe does not pose limits. In cases in which researchers conclude that knowledge simulation outputs are excessive and there are limitations to describe them properly so they can be reused, we recommend that researchers still publish example output files along with products such as visualizations, tables, graphs showing the results. Note that publishing graphs without at least example files is not enough to facilitate a clear rendition of what the outputs that generate those graphs are.

In publishing outputs researchers should consider the following:

* At the moment, DesignSafe does not pose a limitation on the number of files that users can upload and publish.
* Publishing large numbers of outputs entails describing them and providing post-processing code to read them within different applications so that users understand their function.
* Using the description field in the Output category, provide an overview of the results of the simulation. Further details should be included in the Data Report (e.g., format, definition, units).
* Tagging output files is strongly recommended. Output files can be tagged using the list of terms available in the curation interface or by creating your own tags.
* Include code and or directions on how to read output files.
* Some researchers make outputs readable as MATLAB files, or readable by Paraview, ArcGIS and other applications. Instructions on how files were processed should be included.
* If researchers are publishing example outputs, explain how examples were selected and identify their function. Consider including example outputs used producing figures and visualizations published under the Analysis category.

### Analysis

This is a recommended category to inform about post-processing and research methods conducted to analyze the outputs of the simulation. Specifically for knowledge production datasets, we encourage researchers to publish graphs and visualizations that demonstrate the results.

* Describe the goal of the analysis.
* Describe the analysis method used to produce graphs and visualizations.
* List software used in the analysis including version number and reference those in the Referenced Data and Software field.
* For platforms such as Python, provide YAML files for recreating the computing environment.
* Provide detailed information on the methodology used in the analysis or add references to the paper describing the methodology in the Related Work field by choosing the "Context" tag.
* Indicate the provenance of the files used in the analyses by referring back to the Output category or by including the output files in the Analysis category if only some were used.
* As applicable, comment on the limitations of the method/dataset.

### Report

A Report is a required category to publish a simulation dataset. In DesignSafe, the Report is displayed above other categories so that users can quickly find important information about the simulation. Considering reproducibility, the simulation report is a guide for other researchers to understand, reproduce the simulation, and reuse its data for subsequent research. Thus, the report is a more detailed account of the research process than what is feasible to describe in the metadata fields provided by the interface. At the same time, the report is not the paper produced to convey a research project. While we encourage the citation of papers about the research, or even publishing a preprint of the work within the data package, the report focuses on the methods, logic, and tools involved in the data production process required for others to use the data.

The report may entail different pieces of documentation such as readme files for code, a data dictionary explaining the meaning of the variables included in the results, an explanation of the file naming convention, etc. or it may include all the pieces in one document. Below we include suggested sections and content for a report

* Simulation Methodology/Model
  + Provide context by including a brief overview of the research questions.
  + Describe how the simulation runs were structured (e.g. workflow of the simulation).
  + Include a visualization/scheme of the workflow.
  + Include information about spatial and temporal scales and resolution.
  + Note the typical runtime of the workflow and the system configuration used for simulations.
  + Include any problems that you anticipate the end user will face. Example entries include common model errors, their reasons, and how to address them.
* Data sources/input
  + Describe the inputs including information about the data sources and any relevant characteristics such as size, resolution, and units.
  + If the data is tabular, include a data dictionary to describe the column contents.
* Inputs pre-processing.
  + Describe pre-processing steps.
  + Include scripts or queries used to collect input data.
* Outputs
  + Is this a data production or a knowledge production simulation?
  + Explain if all outputs are included in the dataset or only examples.
  + Explain if this is the raw data or if it has undergone post-processing.
  + Describe the data post processing.
  + If the data is tabular, include a data dictionary to describe the column contents.
* Description of results.
  + Provide a summary of the results obtained.
* Analysis.
  + Describe analysis methods and results.
* Include readme for code/s included in the publication.
  + Mention how to seek help related to this data or method.
* Guidelines for data reuse.
  + Mention possible applications for the method and/or dataset. Also, mention applications (if any) where the dataset/method should not be used.
  + Describe naming convention used for folders and files for ease of use.
  + Especially for large outputs if you have organized the files in subfolders, the content and way in which files are organized has to be described to allow others to locate precise files they may need.
  + If available, provide alternative sources of input data that might facilitate the transferability of the method to a new study region or application.
  + Mention how to seek help related to this data or method.
  + Point to information on how to reuse data in the Data Depot workspace or to bulk download data available in the DesignSafe Help section.

Below are examples of simulation publications in DesignSafe that include Reports:

Kameshwar, S., N. Vishnu, J. Padgett. (2019) "Earthquake analyses for portfolios of seven highway bridge classes", in *Response and Fragility Modeling of Aging Bridges Subjected to Earthquakes and Truck Loads*. DesignSafe-CI.<https://doi.org/10.17603/ds2-7rdw-cw70> v1

Asher, T., (2019) "Hurricane Matthew Storm Surge and Wave Simulations", in *Hurricane Matthew Storm Surge and Wave Simulations with Data Assimilation*. DesignSafe-CI. <https://doi.org/10.17603/ds2-ne6j-s897> v1

# Other Best Practices

### Ancillary **Code**

Researchers may include scripts within different parts of a simulation to run the model, as a mature model, to pre-process the inputs, to process the outputs, and to generate graphs and visualizations of the results. These resources may be already published elsewhere, may be ready to publish as research software in DesignSafe, or may be scripts used in this particular simulation but do not amount to a software publication.

To publish code in conjunction with data, we recommend to:

* Use a consistent code style and widely accepted conventions and design patterns. Using standard style guides (such as PEP 8 <https://peps.python.org/pep-0008/> for Python) might facilitate interpretation.
* Use Docstrings (i.e., documentation string) to document functions, modules, classes, and packages. Use standard Docstring format (e.g., NumPy/SciPy or StructuredText).
* Use comments in the code file to document the workflow and facilitate reuse.
* Share the code used for the simulation on DesignSafe. Codes related to data collection, data processing, and visualization should be included within the simulation model categories in which the code was used.
* Create an online version of the code via hosting services such as GitHub (https://github.com/). This allows users to use the latest version of the code, with potential bug fixes, if they prefer to do so.
* If the code/software has been published, include the information in the Referenced Data Software field.
* Provide information on how to cite the code in the Report. An example BibTeX entry is shown below. Alternatively, use the built-in citation support available on GitHub (https://citation-file-format.github.io/).

### File Formats

Due to the diversity of data and research methods used by our community, there are no current restrictions on the file formats users can upload to DesignSafe. However, for long-term preservation and interoperability purposes, we recommend and promote storing and  publishing data in [open formats](https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-formats), and we follow the [Library of Congress Recommended Formats](https://www.loc.gov/preservation/resources/rfs/TOC.html). A list of accepted and recommended file formats in DesignSafe is available [here](https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/best-practices/).

### File Naming Conventions

Use short, informative and consistent names of your files and of folders when it applies. For example, a naming convention could be **YY-MM-DD\_Simulation\_Location\_Characteristics**. Where **YY** is the year, **MM** is the month, **DD** is the day, **Simulation** refers to the performed simulation number or model used, **Location** defines a particular component or structure ID, and **Characteristics** is any other required feature or parameter considered for this simulation. The file naming convention used should be described in the Data Report.

### Referencing Simulation Datasets

Any paper published referencing the simulation dataset should be included as a citation in the Related Work field using the “Cited by” tag.

Related Works can be added any time after the dataset is published by following the Amends and Version pipeline within a Project. We recommend that users take the time to include such references as they will be reflected in the citation count on the landing page.

Chen, Q., C. Johnson, S. Miao. (2022) "Delft3D/XBeach Model", in *Morphodynamic modeling of hurricane impact on Louisiana low-lying coast using Delft3D and XBeach*. DesignSafe-CI. <https://doi.org/10.17603/ds2-cks2-9e45> v1

**References**

1. Gretchen L. Mullendore, Matthew S. Mayernik, Douglas C. Shuster. “Open Science Expectations for Simulation-Based Research.” *Frontiers in Climate*, vol. 3, Nov. 2021, p. 763420 , <https://doi.org/10.3389/fclim.2021.763420>.
2. Maegen B. Simmonds, William J. Riley, Shreyas Cholia, Charuleka Vardharajan. *Addressing Model Data Archiving Needs for the Department of Energy’s Environmental Systems Science Community*. May 2020. *eartharxiv.org*, <https://eartharxiv.org/repository/view/260/>.
3. Maria Esteva, Craig Jansen, Pedro Arduino, Mahyar Sharifi-Mood, Clint N. Dawson. “Curation and Publication of Simulation Data in DesignSafe, a Natural Hazards Engineering Open Platform and Repository.” *Publications, Special Issue “Selected Papers from Open Repositories 2018”*, vol. 7, no. 3, 3, Sept. 2019, p. 51. <https://doi.org/10.3390/publications7030051>.

1. Beyond the specific functionalities available these recommendations can be used by researchers publishing simulation data in other platforms. [↑](#footnote-ref-0)