ICSU CODATA
ICSU WORLD DATA SYSTEM
SciDataCon 2016

Conference paper

Modelling Natural Hazards Engineering Data to Cyberinfrastructure

Summary

DesignSafe-CI is an end-to-end data lifecycle management, analysis, and publication cloud platform for natural hazards engineering. To facilitate ongoing data curation and sharing in a cloud environment that is intuitive to the end users, developers and curators teamed with experts in the different hazards to design data models and vocabularies that map their research workflows and domain terminology. The experimental data models - six - emphasize provenance through relationships between research processes, data and their documentation, and highlight commonalities between experiment types. They mediate between the user interface and the repository layers of the cyberinfrastructure to automate tasks such as organizing data and facilitating its description. Using data from triaxial experiments, we conducted a user evaluation of the geotechnical data model, both for its fitness to real data and for purposes of data understandibility during reuse. The results of the evaluation guided testing and selection of the Fedora 4 repository backend to enhance data discovery and reuse.

Introduction

DesignSafe-CI (designsafe-ci.org) is an NSF-funded cloud-based research environment that facilitates the full lifecycle of data required by scientists and practitioners to effectively address the threats posed to civil infrastructure by natural hazards. Researchers in the community conduct small and large scale experiments in Experimental Facilities (EFs), each of which has distinct equipment to simulate a hazard - tsunamis, earthquakes, wind - and measure its impact on physical and on hybrid models.[1] Researchers can upload data to the cyberinfrastructure (CI) from the experimental project's inception, and carry the data through subsequent analyses, curation, publication, and reuse stages. But to adopt the CI, users require tools and functions that represent their research practices.

From the beginning of the DesignSafe-CI project, a team of experts in the different hazards and the staff of the EFs were guided to provide structured information about their research practices and requirements to move their research into the cloud. Through interviews, site visits to EFs, and workshops, we learned that while researchers spend extensive time documenting their experiments and organizing their data, archiving and publishing are left to students to accomplish, and to curators, that may not understand the core research, to validate. Across the different types of experiments, the common thread was the complex structure and variety of the datasets, which include design drawings, sensor lists, and outputs from multiple sensors and experiment runs. As well, the community lacks metadata standards to organize and

[1] To find information about the different Experimental Facilities and types of experiments conducted go to https://www.designsafe-ci.org/facilities/experimental/

communicate the meaning of their data. Data discovery is also key to the community. For understandability and ease of reuse, data has to be retrieved in relation to information about the sensors from which it generates as well as to contextual documentation.

To work on solutions, developers, curators and researchers embarked on a joint process of modelling research workflows and data to create, in the cloud, an environment that reflects the methods, terminology, and tools used by the community. This resulted in a general experimental data model to which special terms (as metadata) representing each of six different experiment types can be imposed to suit data curation, representation, and access needs.

Modelling Geotechnical Experimental Data and Processes

Researchers were asked to describe and draw their research workflows from the moment in which they begin the experiment design up to data publication, including processes that are accomplished in the lab and those that can be completed in the CI such as: data analysis, organization, description, and archiving. They noted the types of files generated, the software tools used to process and analyze data, and the documentation that is indispensable for proper data interpretation and reuse. After analyzing the workflows of all experiment types, and identifying common processes as entities as well as the relationships between those, the curators derived a general experimental data model. Entities are the main groupings into which data and documentation are organized; they are: *Project, Model_Configuration, Event, Sensor, and Analysis*. They may have different - one to many or many to many - relationships between them, depending on the configuration, number of models, events, and analyses of each project. Following, the experts identified and defined specialized terms[2] that will allow customizing the general data model per experiment type.[3] Terms can correspond to one or more entities through semantic relations (e.g. is part of, is output of, etc.) so that when users select them as tags in the GUI, the associated objects are automatically organized and connected through metadata.

Data Model Evaluation

The experimental data model was evaluated to understand how it fits real world datasets, observe how users use it to organize and describe their data, find out if they are understood by others trying to reuse data, and to inform the design of the GUI and the development of data discovery mechanisms. A total of eight triaxial experiments were performed at the geotechnical engineering research laboratory at University of California, Los Angeles. Experimental measurements from the tests are post-processed to obtain the mechanical properties of loose sand samples. This dataset was deemed useful to evaluate fitness with the model for geotechnical experiments. Within the data model, an *Event* is defined as a loading condition imposed on a *Model.* Each *Event* produces an output data file, and these files are then analyzed to obtain soil properties. Each *Event* has an associated *Sensor List* containing metadata regarding the sensors and data acquisition channels used in the events.

[2] The terms and definitions are entered in YAMZ at: http://www.yamz.net Search under the hashtag #DesignSafeExperiment.

[3] We note that in the process of normalizing the vocabularies we found that many terms are common across experiments types.

For evaluation we first asked researchers to store and organize the dataset in Box.com. Box was selected for the testbed because it allows describing and tagging folders and files which can be shared. Besides from the data, extensive documentation about the experiment, a data dictionary, and post-processing tools were included. The files were organized in folders based on the data model, and tagged using the geotechnical vocabulary to clarify the nature of each object. We asked researchers to suggest changes to the terminology when necessary. In addition, because Box is a hierarchical file system, researchers indicated how groups of files relate to each other through complicated file naming schemas. Observing the way in which researchers interpreted the model allowed adjusting the relationship between entities and revisiting the meaning of some terms. For example, they suggested using the name Specimen and Experiment Planning instead of Model Configuration. Figure 1 below shows how this dataset corresponds to the geotechnical experimental data model.

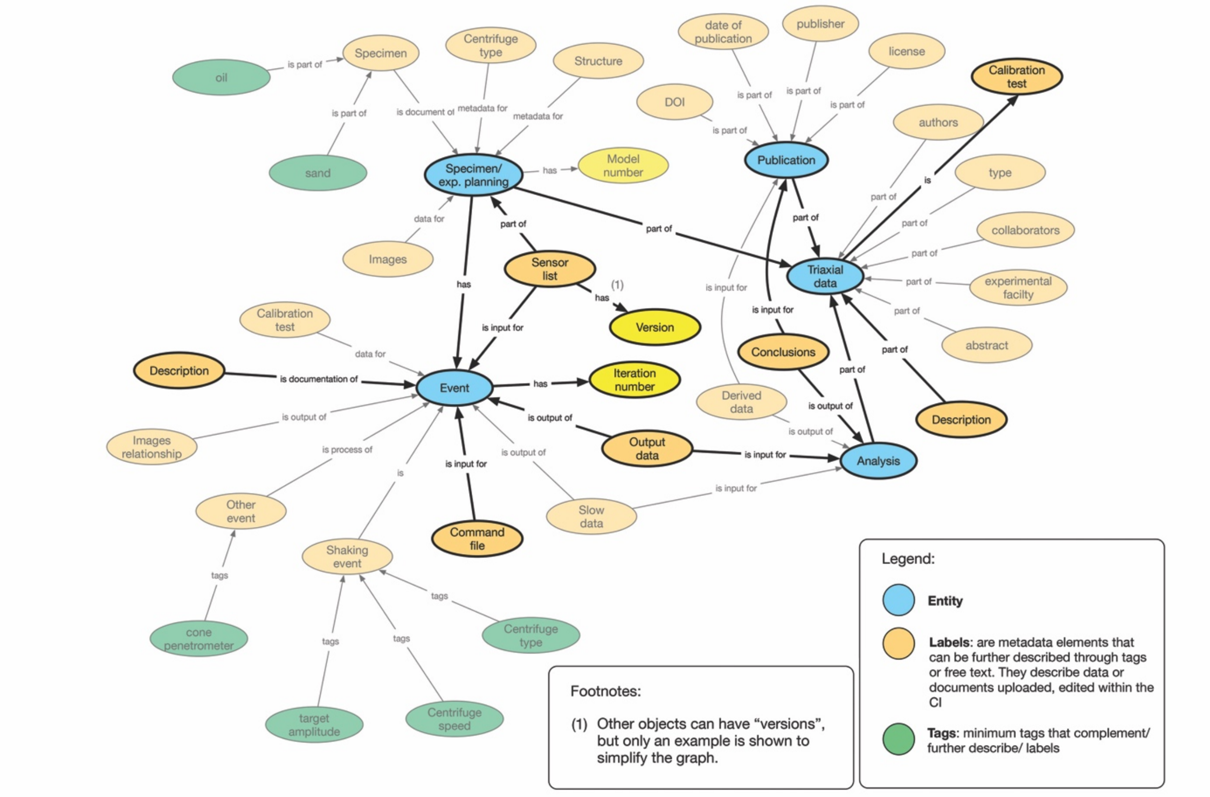

Figure 1: Triaxial dataset entities and metadata terms (highlighted) are superimposed to the geotechnical experiments data model (transparent) to evaluate its fitness.

To evaluate understandibility of the data organization, twenty-five students taking the Advanced Geotechnical Design course completed an assignment in which they were required to download the experimental data and use it to compute the mechanical properties of the sand specimens. A pre-assignment survey assessed their prior experience with data, revealing that 96% had previously re-used data produced by others. A post-assignment survey containing 12 questions with Likert-type scale responses assessed the ease with which they were able to reuse the triaxial data in relation to its organization, documentation and description.

Important observations from this exercise are: (1) a strong majority of students understood the way the data was organized and tagged, (2) most students found the documentation about the data to be useful, and (3) students were dissatisfied with the extensive click through required to download all the data and files to complete the assignment. The primary conclusions are that comprehensive data organization and documentation are key to its understandibility, and that access mechanisms allowing precise retrieval of the data in relation to corresponding contextual documentation would ease reuse. Also, the team identified redundancies in the storage of data that contributed to the negative feedback from the students (e.g. the same sensor list was stored repeatedly as it was used in many events) and made necessary changes to the model.

Mapping to a Repository Backend

Following, the team explored backend repository solutions that would align with the gathered feedback. Fedora 4[4] was selected for evaluation primarily because of its native RDF capabilities, which enable structuring metadata to reflect the flexible, often non-hierarchical relationships between the entities. To submit the triaxial data according to the reviewed data model, we mapped the entities to the repository's own data model, which is based on parent/child relationships. Files were uploaded and related through DublinCore relation type attributes isPartOf and HasPart. In this way, all the *Experiment Planning* files were related to the data files corresponding to each *Event (8), and* each *Event* to one of two *Sensor Lists* used in the experiments*.* The relationships were then queried using SPARQL to retrieve, for example, all of the events conducted with one sensor list and vice-versa.

We concluded that Fedora will facilitate searching and will better support the design of a GUI that provides direct access to the data and its documentation. As well, Fedora 4's automatic generation of UUIDs will enable the system to refer to objects in relationships unambiguously, in contrast to a dependence on file naming seen in the Box testbed. Importantly, Fedora 4 also supports the long-term integrity of data through built-in, automated digital preservation features including checksum generation and validation, versioning, and robust audit logs of all data objects. The RDF functionality preserves data provenance by formally representing data component relationships through persistent, structured, system-embedded metadata.

Conclusions

This work highlights the importance of involving users early in the design of cyberinfrastructure for research. Modelling was based on varied research workflows, intended to map methods and tools that are familiar to this community to cyberinfrastructure. The model was evaluated by users that needed to reuse data and modifications were made accordingly. Fedora 4 was selected as a good fit for a research data repository of enduring importance to civil infrastructure. Further iterations in the CI development will continue involving data creators and users in the process.

Acknowledgements

This research was supported by a grant from the National Science Foundation (#1520817).

Competing Interests

The authors declare that they have no competing interests

[4] Fedora 4, http://fedorarepository.org/documentation