IJDC | *General Article*

Designing and Building Interactive Curation Pipelines for Natural Hazards in Engineering Data

Maria Esteva
Texas Advanced Computing Center
University of Texas at Austin

Craig Jansen
Texas Advanced Computing Center
University of Texas at Austin

Josue Balandrano Coronel
Texas Advanced Computing Center
University of Texas at Austin

Abstract

To design data curation pipelines within DesignSafe-CI, we gathered requirements and sought regular guidance from a group of experts in different aspects of natural hazards engineering research. Upon achieving understanding of experimental, simulation, hybrid simulation and field reconnaissance research workflows, we created four data models to guide data organization and developed specialized vocabularies as metadata. We then translated the models and metadata to interface design (front-end), and selected the infrastructure resources that would support curation and publication functions (back-end). We used iterative design and testing, including the use of interactive mockups of the GUI, to communicate and elicit feedback from the experts, and mapped real datasets to the mockups to evaluate the fitness of the data models, the clarity of the curation tasks. To address the problem of big data interfaces, we provide data representations that highlight the structure of the datasets and the possibility to browse their components in relation to provenance.

Received 22 January 2018 ~ *Accepted* 22 January 2018

Correspondence should be addressed to Maria Esteva, 3206 South Oak Drive, Austin, TX, USA. Email: maria@tacc.utexas.edu

An earlier version of this paper was presented at the 13th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

To study the characteristics and the impact of natural hazards and develop methods to prevent damages to populations, structures, and the environment, engineers employ diverse research methods including: experiments, simulations, hybrid simulations, and field reconnaissance (field recon). Each may generate multiple and different types of data that can be modified across research steps, resulting in large, and multi-relational datasets. Due to their sizes, scientific complexity, and relational structure, curating these datasets so that they are understandable and easy to reuse is challenging. There are no community standards to curate natural hazards engineering data and, while there is previous work on curation of large-scale experiments (Pejša et al., 2014), there are few isolated examples of large simulation (GRIIDC)[1], and field recon (GEER)[2] data publications to draw from.

We understand curation pipelines as the front-end graphical user interfaces (GUI) to organize, describe, verify, and publish different natural hazards datasets, and the back-end infrastructure that supports these functionalities along with the formation of standardized metadata, and the long-term preservation of the data. In designing the pipelines within the end-to-end data management and analysis platform DesignSafe-CI (DS-CI) (DesignSafe-CI)[3], the curation and web development team's goal was to model how researchers conceive their investigative workflows in order to integrate curation to the data analysis tasks conducted in the CI including data transitions between active study, in-curation, and static lifecycle stages. To spike adoption, curation tasks had to be relatively easy to undertake and had to simplify documentation, as researchers' lack of engagement with the curation process is a known problem (Borgman et al., 2016; Scaramozzino et al., 2012). Importantly, the published data had to convey the sophistication of each research project in ways that other users could understand. For this we had to address the problem of big data interfaces, typically represented in open repositories as interminable lists of files with high level descriptive metadata, which present difficulties to navigate and understand. Lastly, we had to decide the infrastructure components that would support all the functionalities.

From the DS-CI's project's inception, experts in the different research methods were involved in the design and testing of the pipelines (Rathje et al., 2017). To capture their knowledge and transfer it as interactive curation steps we followed a methodological approach and employed iterative design and testing of the curation interfaces. This paper focuses on the process that we followed to understand natural hazards engineering research and gather community requirements. How we modelled the researcher's knowledge and feedback as data models for data organization and as metadata for description, and how we conceived and evaluated curation tasks in a GUI. Finally, we describe the back-end infrastructure architecture.

[1] GRIIDC: https://data.gulfresearchinitiative.org/data/R1.x134.114:0008

[2] GEER: http://www.geerassociation.org/

[3] DesignSafe-CI: https://www.designsafe-ci.org

Modelling Natural Hazards Engineering Research

DS-CI[4] is an end-to-end data management, analysis, and publication platform funded by the National Science Foundation (2015 to 2020). It is one of the components of the Natural Hazards Engineering Research Infrastructure (NHERI) to improve resilience and sustainability of infrastructure and critical lifelines. DS-CI supports large-scale analysis and publication of data generated during experiments, field recon, simulations, and hybrid simulations research projects. It provides open access to large-scale computational resources and software tools, facilitates curation and constitutes an open repository for data. To the development and curation team, the first step was to understand the characteristics of natural hazards engineering research. Realizing its complexity set the tone to our work. It revealed the needs to involve domain experts and to devise new ways to address the data diversity and scale.

Natural Hazards Engineering Research Community Involvement

To attain a foundation from which to begin the design, during the first eight months of the project the data curator along with principal investigators and developers, travelled to six experimental facilities (EF) across the country. They observed equipment and methods used to gather data as experiments are conducted, and spoke with the staff about data-keeping and transfer to (see Figure 1). In addition, data curation and publication were discussed in relation to future goals and priorities for the entire platform during two initial user community workshops.

A regular venue to gather information has the form of two requirements teams: simulation and data. Each has five experts that study natural hazards phenomena from a different angle (e.g. geotechnical, wind, storm-surge, structural engineering, etc.) and using different research methods and equipment. The teams meet virtually with the curators and developers every other week to discuss data agenda items and assess progress. The continuous discussions have worked to the advantage of the entire group. As much as the curators need to learn about natural hazards engineering, the domain researchers need to understand what is entailed in data curation and publication. In turn, web developers, who are experienced builders of large-scale data portals, need to learn about digital library and archives tools and methods to produce FAIR data[5]. During the visits, workshops, and regular meetings we also learned what the community perceived about and wanted from the curation process. All this information became the foundation to create the data models that guide the organization of the datasets, and the metadata to describe them. It also guided us through the GUI and the architecture infrastructure design.

[4] NHERI: https://www.designsafe-ci.org/about/

[5] FAIR Data Principles: https://www.force11.org/group/fairgroup/fairprinciples

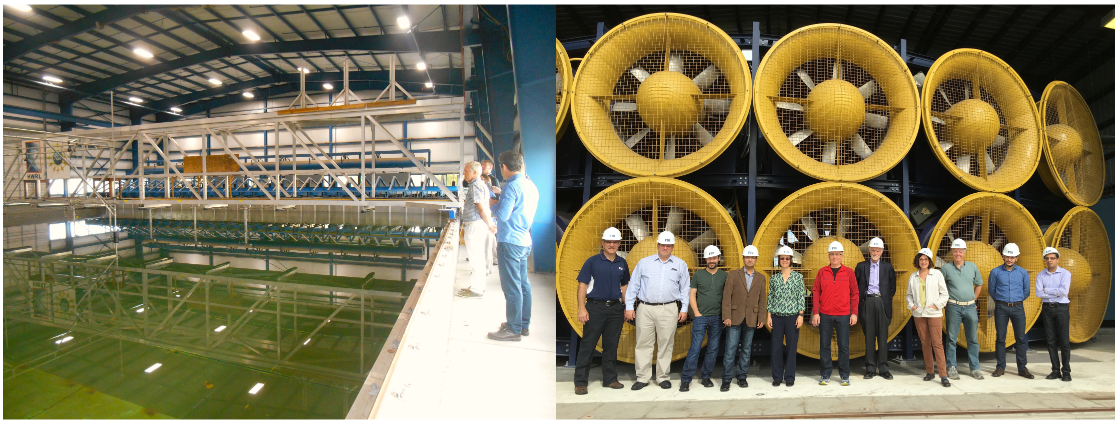


Figure 1. To the left the wave basin at the University of Oregon EF. To the right, the Wall of Wind at Miami International University EF. Each experiment involves the preparation of a physical model with sensors that measure the loads that are emitted from the wind and wave sources. The equipment highlights the scale and complexity of the experiments.

Characteristics of Natural Hazards Research and Datasets

Below is a selection of the main characteristics and requirements gathered from the experts. Most issues have been addressed in the curation pipelines and the remaining ones are on the works.

List of data characteristics and requirements

1. All research methods can potentially generate thousands of very large files.
2. Most experimental and simulation projects are run more than once. In each iteration, a moving part changes the resultant data.
3. In experimental projects, the configuration of the iterations is unique.
    a) One large-scale experimental project may consist of many individual experiments that are undertaken by different authors, and each experiment may entail several runs.
    b) Each researcher may conceive iterations differently. To some they are runs within an experiment, to others each is an individual experiment.
4. Large-scale experimental projects may take up to one year of preparation and many more to process and study the data. Studying the resultant experiments may take several years in which each is published at a different time.
5. To researchers, the boundaries between active and published data, and between data management, analysis, and curation are blurry.
    a) As researchers conduct analyses, data has to be available for reuse independently of whether it has been published, as it may be the input for a new study and thus a different publication within the project.

6. Large-scale natural hazards research is scientifically complex. To reuse the data, users will have to delve into the details of the projects.
    a) Experimentalists and field recon researchers produce detailed reports.
    b) Experiment reports may take years to compile as many are dissertations.
    c) Simulation researchers do not have agreed-upon documentation practices.
7. Most researchers did not agree on what to preserve and what to discard as by-products of their large-scale experiments and simulations.
8. The majority of researchers did not have a clear notion of what curation entailed, nor how to make their datasets understandable and reusable.
    a) Licensing or the functions of a DOI were unclear to most.
9. There is no standardized metadata to describe natural hazards engineering research data. Some isolated vocabularies are being developed by study groups.
10. Of main concern to the researchers is whether their data has been cited.
11. There was a strong demand for easy, intuitive, streamlined curation tasks.

Modelling Research Workflows as Data Models and Metadata

To capture the researchers' knowledge during the first year of the project we followed a structured methodology. After a brief explanation of how to express their research workflows, each of the requirements team members and staff from the EFs had to draw or write down the steps, processes, tools, documentation objects, and data products. We also asked them to include the terms that they use to name processes, tools, and resultant data. In addition, we conducted interviews during which the researchers narrated their workflows so we could better capture the processes and their relations. From this information we derived four data models and specialized vocabularies that are used to design and architect the curation pipelines (Figures 2, 3 and 4).

Data models are abstractions. In DS-CI curation pipelines, their role is to represent the main research processes as categories around which data and documentation files can be organized. We created four data models for: simulation, experiments, hybrid simulations, and field recon research types (Esteva et al., 2016). Figure 2 below shows a simulation workflow drawn by a researcher, and a corresponding data model showing the relations between the main processes/categories – of a storm surge simulation.

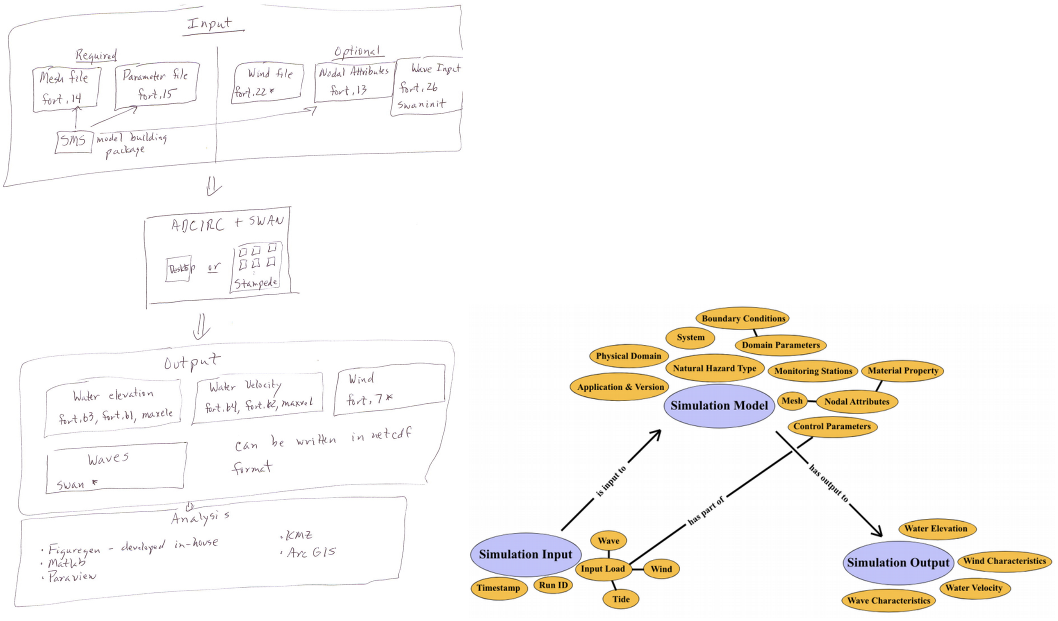


Figure 2. To the left the workflow for storm surge simulation drawn by a researcher. To the right a section of the corresponding data model for simulations with the main categories (purple) coupled with specialized terms (yellow) that describe storm-surge simulation research.

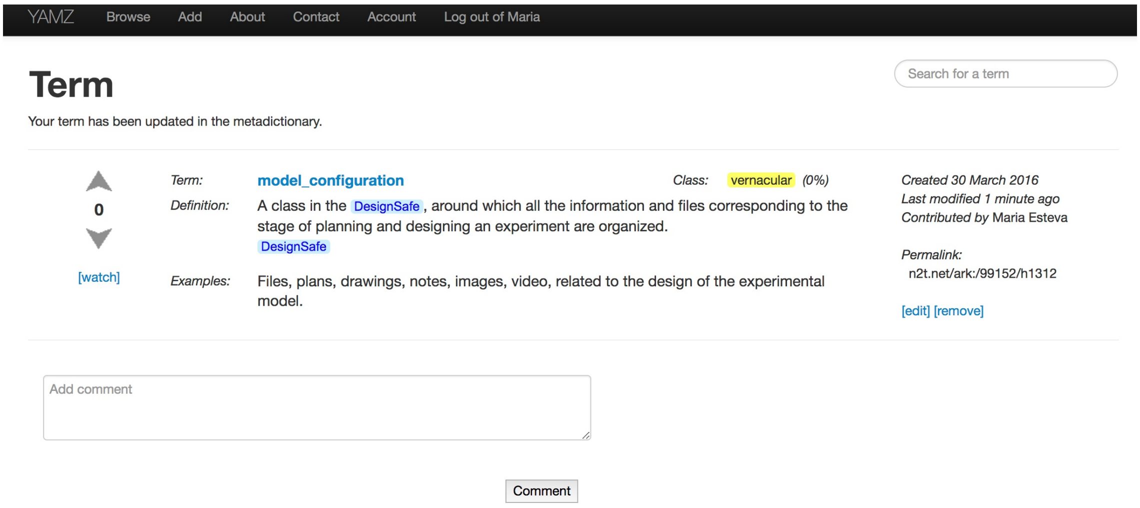
Table 1. Data model categories for each research method curation pipeline in DS-CI.

Research Method	Organizing Categories
Experiments	Model configuration, sensor information, events, data analysis, report
Simulation	Simulation input, simulation model, simulation output, data analysis, report
Hybrid Simulation	Global model, simulation coordinator, sensor information, analytical substructure, physical substructure, analysis, report
Field Reconnaissance	Site, observation, analysis, report

Table 1 above shows the categories for each data model. The labels were agreed upon by the requirement teams to normalize semantic differences across terminology used by researchers using similar methods. With the labels and specialized terms contributed by the experts we produced vocabularies to describe data according to different study approaches: structural, wind storm-surge, structural, wave-basin, geotechnical, etc. The definitions were also written by researchers. Figure 3 shows the term *model configuration* recorded in the online meta-dictionary (YAMZ)[6]. By introducing the main processes and their relations, the data models represent the structure and provenance of the data in connection to research steps. In the GUI, the

[6] YAMZ: http://www.yamz.net

vocabularies are metadata elements for purposes of aiding a research project's documentation and minimizing manual entry.



Term

Your term has been updated in the metadictionary.

Term: model_configuration

Class: vernacular (0%)

Definition: A class in the DesignSafe, around which all the information and files corresponding to the stage of planning and designing an experiment are organized.
DesignSafe

Examples: Files, plans, drawings, notes, images, video, related to the design of the experimental model.

Created 30 March 2016
Last modified 1 minute ago
Contributed by Maria Esteva

Permalink:
n2t.net/ark:/99152/h1312

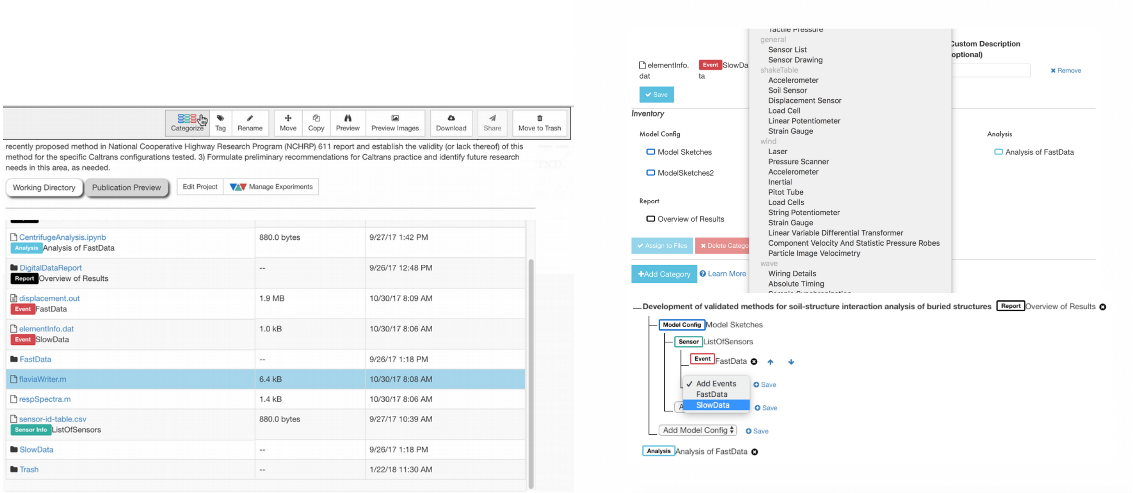
Figure 3. Snapshot of the definition for the term model_configuration in YAMZ.

Curation Pipelines Front-End: Interfaces

The interface design transforms the data models and metadata into interactive tasks that enact data organization, description, and publication activities.

Transitioning Between Active Research, Curation, and Publication Stages

To analyze, curate, and publish their data in relation to the rest of the DS-CI platform, users need private and shared workspaces to manage active data as well as outlets for public data instances. All of this happens in the Data Depot, where users can store and access data individually in My Data, and create shared projects and access existing ones in My Projects. Transitions between active research, curation, and publication are feasible within a project. From the working directory, users can upload, copy and share data; select it for computational analysis; conduct curation tasks progressively; and track and reuse data already curated and or published (Figure 4).



recently proposed method in National Cooperative Highway Research Program (NCHRP) 611 report and establish the validity (or lack thereof) of this method for the specific Caltrans configurations tested. 3) Formulate preliminary recommendations for Caltrans practice and identify future research needs in this area, as needed.

CentrifugeAnalysis.ipynb Analysis Analysis of FastData	880.0 bytes	9/27/17 1:42 PM
DigitalDataReport Report Overview of Results	--	9/26/17 12:48 PM
displacement.out Event FastData	1.9 MB	10/30/17 8:09 AM
elementInfo.dat Event SlowData	1.0 kB	10/30/17 8:06 AM
FastData	--	9/26/17 1:18 PM
flaviaWriter.m	6.4 kB	10/30/17 8:08 AM
respSpectra.m	1.4 kB	10/30/17 8:06 AM
sensor-id-table.csv Sensor Info ListOfSensors	880.0 bytes	9/27/17 10:39 AM
SlowData	--	9/26/17 1:18 PM
Trash	--	1/22/18 11:30 AM

Figure 4. To the left a snapshot of the working directory showing files/directories: selected for curation, curated with colored tags and metadata, and not. To the right, the GUI to assign categories to files/directories, describe them with specialized vocabularies, and relate categories.

Modelling Curation in the GUI

Once users create a project and at any point in the research lifecycle (ideally early on), they can choose to select a research type as experiment, simulation, hybrid simulation or field recon and start curation. Due to the flexibility to create any number of research methods instances, those will be tied together at the project level. In the interface we operationalized curation as a two-stage process, each involving tasks: a) categorization and description, and b) publication. In the first stage users: 1) select files/directories from the working directory, 2) categorize them as corresponding to one or more categories, 3) describe them using the specialized vocabulary, and 4) relate categories (See Figure 5). The publication stage involves: 1) reviewing selected files 2) verifying metadata, and 3) choosing licenses and signing the repository agreement. Once the publication package is submitted, the project and each research instance obtain DOIs. If users want to publish new experiments or simulations at a later time, those will receive DOIs that will be related to the project through the Data Cite metadata.

Datasets representations

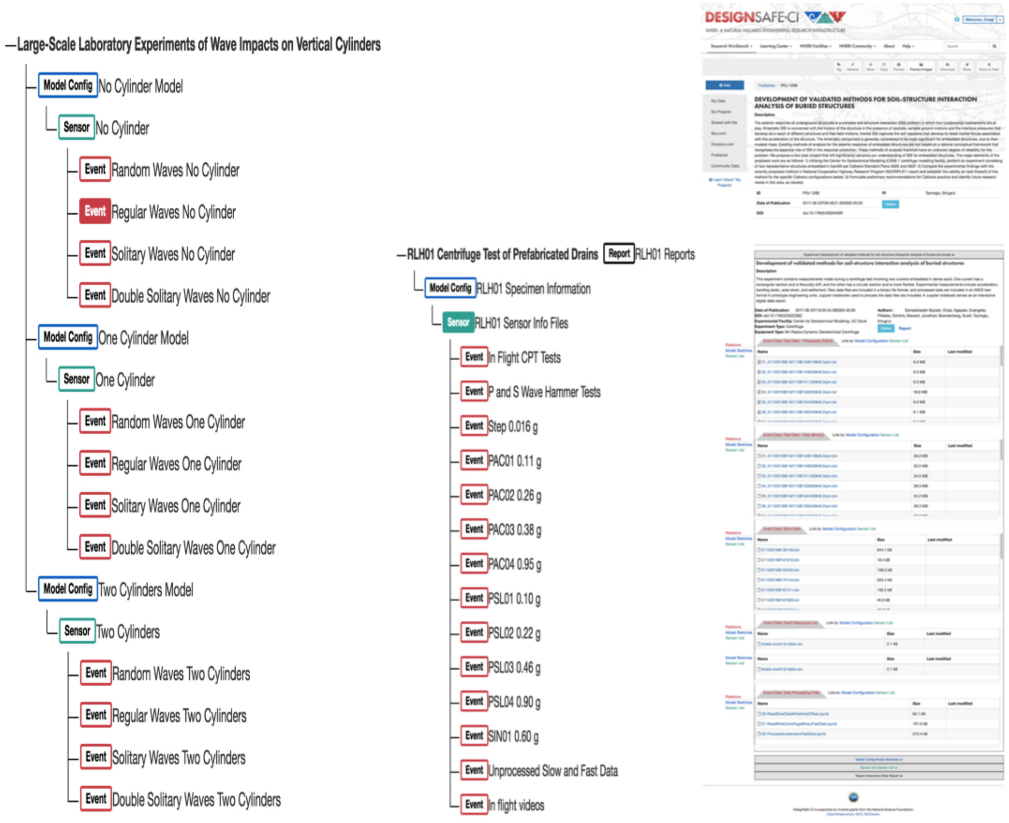

Large-Scale Laboratory Experiments of Wave Impacts on Vertical Cylinders
Model Config No Cylinder Model
Sensor No Cylinder
Event Random Waves No Cylinder
Event Regular Waves No Cylinder
Event Solitary Waves No Cylinder
Event Double Solitary Waves No Cylinder
Model Config One Cylinder Model
Sensor One Cylinder
Event Random Waves One Cylinder
Event Regular Waves One Cylinder
Event Solitary Waves One Cylinder
Event Double Solitary Waves One Cylinder
Model Config Two Cylinders Model
Sensor Two Cylinders
Event Random Waves Two Cylinders
Event Regular Waves Two Cylinders
Event Solitary Waves Two Cylinders
Event Double Solitary Waves Two Cylinders
RLH01 Centrifuge Test of Prefabricated Drains
Report RLH01 Reports
Model Config RLH01 Specimen Information
Sensor RLH01 Sensor Info Files
Event In Flight CPT Tests
Event P and S Wave Hammer Tests
Event Step 0.016 g
Event PAC01 0.11 g
Event PAC02 0.26 g
Event PAC03 0.38 g
Event PAC04 0.95 g
Event PSL01 0.10 g
Event PSL02 0.22 g
Event PSL03 0.46 g
Event PSL04 0.90 g
Event SIN01 0.60 g
Event Unprocessed Slow and Fast Data
Event In flight videos
DESIGNSAFE-CI
DEVELOPMENT OF VALIDATED METHODS FOR SOIL-STRUCTURE INTERACTION ANALYSIS OF BURIED STRUCTURES

Figure 5. First to the left a tree from a wave experiment. While different to the second tree belonging to a geotechnical experiment, the main categories work well across both types. The third snapshot corresponds to the browsing interface of the geotechnical dataset.

We represent published datasets as trees and browsing interfaces. Both allow identifying provenance as the processes from which data generates and in relation to categories and to metadata to facilitate data navigation, understandability, and access.

Curation Pipeline Backend: Infrastructure

The backend architecture supports the transition between active to published and preserved data (See Figure 6). Active data is stored in Corral, a geographically replicated high performance storage resource (TACC)[7]. As users select files, assign them to categories, and label them with specialized terms and written descriptions, the metadata about their research project is forming through the AGAVE API (Dooley et al., 2018) which manages active data ingestions, deletions, and transfers. Published data is sent to Fedora 4 (DURASPACE)[8], which provides preservation functions and standardizes the metadata. Upon sending the verified publication package to Fedora, the

[7] Corral High Performance and Data Management: https://www.tacc.utexas.edu/systems/corral

[8] Fedora 4: https://wiki.duraspace.org/display/FEDORA4x/Fedora+4.x+Documentation

metadata is mapped to the Prov and DC schemas for exchange and discoverability. Integration between DS-CI built in Django and all services is realized through Restful API calls.

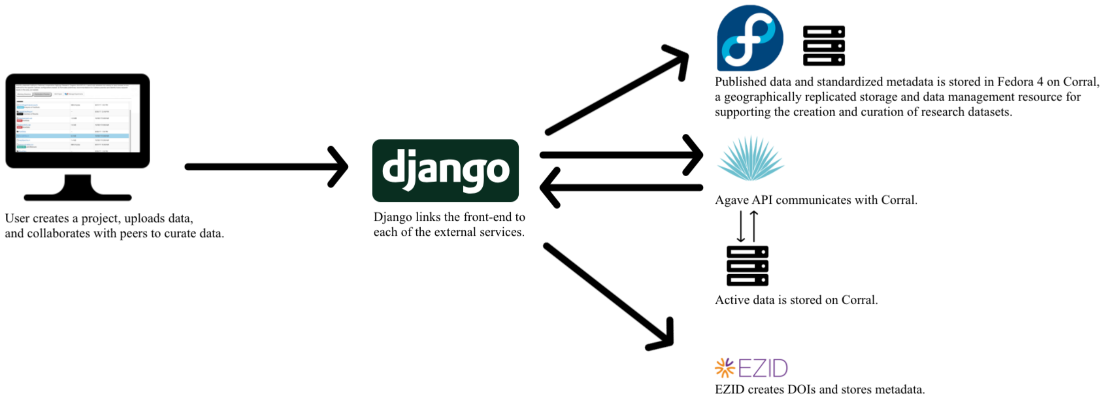

User creates a project, uploads data, and collaborates with peers to curate data.
django
Django links the front-end to each of the external services.
Published data and standardized metadata is stored in Fedora 4 on Corral, a geographically replicated storage and data management resource for supporting the creation and curation of research datasets.
Agave API communicates with Corral.
Active data is stored on Corral.
EZID
EZID creates DOIs and stores metadata.

Figure 6. Infrastructure architecture of DesignSafe-CI curation pipelines.

Evaluating the Pipelines

We used iterative design to evaluate and refine the pipelines. Amongst the many issues to focus, two important ones were: learning if the interactive curation processes mapped the users' conceptions of their research workflows, and if the final representation improved the understandability of the datasets. While researchers had no problem conveying their workflows, obtaining their feedback to model curation activities to a GUI was at times frustrating. Most of them had not considered systematic ways of curating their data beyond using file-naming conventions and hierarchical folders so they could not tell us activities or steps they wanted to follow. On our side, we had difficulties communicating curation concepts and goals. Improvement in communication was achieved through interactive mock-ups prepared to illustrate curation steps. Over the mock-ups, researchers expressed doubts, realized the adequacy of the metadata, added and removed features, and changed their order and placement. They could then "see" what curation implied and imagine how and when they wanted to do it.

We also used the mock-ups to map real datasets to tasks and representations and assess the fitness of the models and interfaces. This has allowed adjusting the design through consensus before major code is written and changes become difficult to implement. In addition, prior to major production releases we ask the researchers to test curation in development mode, and we observe while they interact with their data.

Through compromises and by adjusting terminology (e.g. the term *model_configuration* was intensively discussed by the group), we created data models are generalizable enough to fit datasets from diverse research projects (Esmaeilzadeh et al., 2017; Bernier et al., 2017). A few researchers would like a more prescriptive wizard-style curation GUI, but to the majority, the built-in flexibility provides more freedom to organize their data. About the publication, the experts agree that the representations make it easier for others to understand and reuse the datasets and thus, worth going through the curation pipeline. However, they are still not convinced of the role of the specialized vocabularies. They need to see it in action once the filtered search is developed in DS-CI.

Conclusions

Cyberinfrastructure projects are at the intersection of domain science, big data, computation, and digital libraries and archives best practices. Understanding natural hazards engineering research is a continuous process for data curators and developers, and curation concepts require time to sink in and form part of the researchers' workflows. Scarce curation foundations for natural hazards engineering data demanded to start our curation work developing data models and metadata, and those activities will have to be further undertaken by the community for broader input and standardization. The solutions developed for natural hazards engineering data suggest new paths for progressive, online, curation activities and can be generalized to other domains. They address transitions across research lifecycle stages and big data interfaces. In the next future we plan to continue working on the users' requirements. We will further automate curation tasks, and use standardized metadata to implement search optimization strategies to promote and measure data reuse. So far, we have evaluated our design through the eyes of a dedicated group of experts. As more datasets are curated and published in DS-CI, we will undertake evaluation with new users to evolve an improve through their feedback.

Acknowledgements

DesignSafe-CI is supported by multiple grants from the National Science Foundation. We acknowledge the researchers in the NHERI Experimental Facilities and the members of the Data and Simulation Requirements Teams for their guidance.

References

Bernier, C., Yuxiang, L., Padgett, J., Dawson, C.N., Lomonaco, P., & Cox, D. (2017). *Large-scale laboratory experiments of wave impacts on vertical cylinders*. [Data set]. DesignSafe-CI. doi:10.17603/DS27D4G

Borgman, C.L., Golshan, M.S., Sands, A.E., Wallis, J.C., Cummings, R. L., Darch, P.T., & Randles, B.M. (2016). Data management in the long tail: Science, software and service. *International Journal of Digital Curation, 11*(1), 128-149. doi:10.2218/ijdc.v11i1.428

Dooley, R., Brandt, S., & Fonner, J. (2018). *The Agave platform: An open, science-as-a service platform for digital science.* Proceedings of the Practice and Experience on Advanced Research Computing (PEARC '18). ACM, New York, NY, USA, Article 28, 8 pages. doi:10.1145/3219104.3219129

Esmaeilzadeh Seylabi, E., Agapaki, E., Pitilakis, D., Stewart, J., Brandenberg, S., & Taciroglu, E. (2017). *Development of validated methods for soil-structure interaction analysis of buried structures.* DesignSafe-CI. [Data set]. doi:10.17603/DS2Z38Z

Esteva, M., Brandenburg, S., Eslami, M., Adair, A., & Kulasekaran, S. (2016). *Modelling natural hazards engineering data to cyberinfrastructure.* Proceedings of the SciDataCon 2016. 11-13 September 2016, Denver, Colorado. Texas Scholar Works. doi:10.15781/T2G-B1Z39H

Pejsa, S., Dyke, S., Hacker, T.J. (2014). Building infrastructure for preservation and publication of earthquake engineering research data. *International Journal of Digital Curation, 9*(2). doi:10.2218/ijdc.v9i2.335

Rathje, E., Dawson, C., Padgett, J., Pinelli, J.P., Stanzione, D., Adair, A., Arduino, P., Brandenberg, S., Cockerill, T., Dey, C., Esteva, M., Haan, F.L., Hanlon, M., Kareem, A., Lowes, L., Mock, S., & Mosqueda, G. (2017). A new cyberinfrastructure for natural hazards engineering. *Natural Hazards Review, 18*(3). doi:10.1061/(ASCE)NH.1527-6996.0000246

Scaramozzino, J.M., Ramirez, M.L. & McGaughey, K.J. (2012). A study of faculty attitudes at a teaching centered university. *College and Research Libraries, (73)*4, 349-255. doi:10.5860/crl-255