DesignSafe Data Depot Repository

Notes Before Completing the Application

We have read and understood the notes concerning our application submission.

True

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

CORE TRUSTWORTHY DATA REPOSITORIES REQUIREMENTS

Background & General Guidance

Glossary of Terms

BACKGROUND INFORMATION

Context

R0. Please provide context for your repository.

Repository Type. Select all relevant types from:

Domain or subject-based repository

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:
Accept

Brief Description of Repository

The DesignSafe Data Depot Repository (DDR) is the open data repository of DesignSafe (https://www.designsafe-ci.org/), the cyberinfrastructure component of the Natural Hazards Engineering Research Infrastructure (NHERI). Funded by the US National Science Foundation, NHERI is a distributed, national network that provides the natural hazards community with state-of-the-art research services, facilities, and support. Currently, the DDR is the only platform for curation, publication, and preservation of datasets generated in the course of natural hazards research. Its mission is publicly available at [1]. The DS-DDR is connected to the other DesignSafe cyberinfrastructure components to enable seamless end-to-end management, analysis, publication, and reuse of large datasets. DesignSafe at large and the DDR align with the NHERI mission to provide the natural hazards research community with open access, shared-use scholarship, education, and community resources aimed at supporting civil infrastructure prior to, during, and following natural disasters [2]. The NHERI organizational structure including DesignSafe, and the role of DesignSafe within the network are detailed in [3].

[1] DDR Mission and History https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/policies/
[2] Natural Hazards Engineering Research Infrastructure, Five Year Science Plan, Multi-Hazard Research To Make a More Resilient World, Second Edition.
https://doi.org/10.17603/ds2-4s85-mc54
[3] NHERI organizational structure https://www.designsafe-ci.org/about

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:
Accept

Brief Description of the Repository's Designated Community.

The DesignSafe Data Depot Repository (DDR) serves the global natural hazards research community, including the NHERI network. It is the designated data repository for all the facilities and groups that form part of the NHERI network. Each of these groups have distinct functions, varied research practices, and data curation and publication requirements,

and we work with each to meet their needs and commitments. We also accept data generated in other national and international facilities and organizations. At large, the repository serves the broad audience of natural hazards engineering researchers, students, practitioners, policy makers, and the general public.

The DDR's collections development policy stating data types, sizes, and formats accepted is publicly available [4]. We accept natural hazards engineering and social and behavioural sciences datasets focusing on the study of natural hazards. In the area of engineering our primary focuses are wind, earthquake, and storm surge hazards data generated through simulation, hybrid simulation, experimental, and field research methods. In social and behavioural sciences (SBE), accepted datasets encompass the study of the human dimensions of hazards and disasters. As the field evolves, we have expanded our policy to include datasets related to COVID-19, fire hazards, and sustainable material management. We also accept data reports, publication of Jupyter notebooks, code, scripts, lectures, and learning materials.

Because of the variety of research methods and the innovative and evolving technologies employed by this community, we accept all file formats. We do provide users with ample guidance on best formats for preservation as well as the possibility to publish datasets in both original and converted formats. Due to the sheer size of the datasets generated through large scale research in the space, we do not currently place a cap on the size of the publications. However, we are observing trends in relation to sizes and subsequent data reuse of these products, which will inform if and how we have to implement data size limit policies.

The DDR operates under the leadership of the DesignSafe Management Team (DSMT) who oversees its conceptual and technical development, establishes and updates policies, and evaluates and reviews practices. Ongoing design, development, and operations are accomplished by a repository team. The DDR's governance structure is publicly documented in [5]. In consultation with the NHERI community, the team develops concepts, functionalities, best practices and implements solutions for the DDR. All developments are consulted with and approved by the DSMT who prioritizes them for production. Mechanisms are in place to gather feedback from users and conduct structured evaluation of the repository capabilities. The latter are an integral part of the operations, to make sure that the repository is meeting the community's expectations and needs.

[4] DDR Collections Development Policy
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/policies/collections/
[5] DDR Governance https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/policies/governance/

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:
Accept

Level of Curation Performed. Select all relevant types from:

B. Basic curation – e.g. brief checking; addition of basic metadata or documentation

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:
Accept B.

Comments

While in many cases we provide additional services that would qualify as C or D, our general answer is B.

Due to the diverse, highly specialized, and sheer sizes of the datasets we accept, our curation approach spans different levels. We conceive curation as a partnership between the community as data producers and the repository. The DDR is a self-service repository with interactive curation and automated publishing functionalities to help users publish quality datasets. Throughout the curation interface, data submitters are guided to use data categorization and tagging (including controlled vocabularies) as interactive functions to organize and describe their data in relation to five data models (experimental, hybrid simulation, simulation, field research, and other) that represent the research methods and project types in the space. We also provide onboarding instructions to stimulate users to publish open data formats, to add required documentation, and to guide them on what/how to write metadata entries. During the publication process, the system conducts automated data completeness and required metadata checks. Users are required to include any readme files, data dictionaries or data reports to their publications. Datasets that do not comply with those requirements do not proceed to publication and the user can edit the metadata and add missing files to complete the submission. Upon publication users can correct basic descriptive metadata entries and add new related works to enhance the documentation.

Data curators conduct virtual office hours to meet with users prior to publishing their data, discuss data organization and documentation, and conduct reviews or their future publications. We have a Slack channel and ticketing system to attend users' questions and concerns as they go through the curation and publication processes. If needed, after publication, users can version their datasets by uploading converted files and enhancing documentation.

Certain members of the NHERI network (e.g. experimental facilities, SimCenter, STEER, CONVERGE, etc.) have data curation and validation methods in place for their datasets. These methods are explained in their data reports which they publish along with the validated datasets.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

Insource/Outsource Partners. If applicable, please list them.

The DesignSafe Cyberinfrastructure functions within the Texas Advanced Computing Center (TACC) at the University of Texas at Austin [6]. TACC designs and operates powerful and innovative computing resources to enable research that advances science and technology. DOI services from DataCite [7] are supported through the University of Texas Libraries. All DesignSafe including the DDR capabilities are assessed on a yearly bases by independent consultants. We contract with the user experience consultant Four Kitchens to evaluate the design, usability, and user satisfaction of the DDR curation and publication interactive interfaces.

[6] TACC https://www.tacc.utexas.edu/about/overview
[7] DataCite https://datacite.org/

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:
Accept.

Summary of Significant Changes Since Last Application (if applicable).

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

Other Relevant Information.

DesignSafe Cyberinfrastructure currently holds about half an exabyte of data, of which roughly 30 terabytes is in the published repository (DDR). Publications range up to a maximum of 5TB; seven current published datasets are over 1TB in size, with several more above 100GB. Publications range in scale from a single file to over 2 million files; there are roughly 100M files in the published repository.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

ORGANIZATIONAL INFRASTRUCTURE

1. Mission/Scope

R1. The repository has an explicit mission to provide access to and preserve data in its domain.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept 4.

Response:

The DesignSafe Cyberinfrastructure (https://www.designsafe-ci.org/) is a comprehensive research environment that provides cloud-based tools to manage, analyze, understand, and publish critical research data to study the impacts of natural hazards on the built environment and society. DesignSafe is part of the National Science Foundation-supported Natural Hazards Engineering Research Infrastructure (NHERI). DesignSafe is funded through the NSF award #2022469.

In alignment with the broader mission laid out for both NHERI and DesignSafe in the NHERI Five-Year Science Plan [1] and in the NHERI 2020 Impact Report [2], DesignSafe provides the natural hazards research community with open-access scholarship, training, and community resources aimed at supporting the society and the civil infrastructure prior to, during, and following natural disasters. In this context, DesignSafe provides curation and publication functionalities, and continuous preservation and access to data generated in the course of research through DDR.

Within the DesignSafe Cyberinfrastructure, the Data Depot Repository (DDR) (https://www.designsafe-ci.org/data/browser/public/) is the platform for curation and publication of datasets generated in the course of natural hazards research. It is an open access, self-service, data repository that enables data producers to

safely store, share, curate, and publish research data towards its permanent publication, distribution, and impact evaluation. Through the DDR, data consumers can search for, access, and reuse published data to accelerate research discoveries. Established in 2016, the DDR also preserves all legacy data and metadata from the Network for Earthquake Engineering Simulation (NEES), a NHERI predecessor that began construction in 2000.

Building and maintaining a data repository was one of the requirements of the NSF grant proposal, and a component for which our performance is evaluated every year. Data publication and availability in the DDR provide evidence of the many research activities performed by NHERI. Details on the DDR functions with respect to DesignSafe and NHERI are available in the About DesignSafe page [3]. The DDR mission and history are publicly available in the Curation and Publication Policies documentation [4]. We note that the DDR is only one component of DesignSafe. In this application and throughout these documents, we refer only to the repository (the public collection of published user data), and not user's private project data, working data, or other DesignSafe products outside the scope of the DDR.

[1] NHERI Five-Year Science Plan, https://doi.org/10.17603/ds2-4s85-mc54
[2] NHERI 2020 Impact Report https://doi.org/10.17603/ds2-1f7x-9a52
[3] About DesignSafe https://www.designsafe-ci.org/about/designsafe/
[4] Data Depot Curation and Publication Policies
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/policies/

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

2. Licenses

R2. The repository maintains all applicable licenses covering data access and use and monitors compliance.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept.

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept 4.

Response:

The DDR data licensing policy is publicly available in our documentation [1]. Considering the research outputs generated by the NHERI community, the types of resources accepted in the repository, and how researchers in the community have conveyed how they want to be attributed for their data publications, the DDR provides five licensing options: two for written works, two for data, and one for software. For written works such as papers, presentations, reports and learning objects, we offer the choice between the Creative Commons Attribution (CC-BY 4.0) license, which allows retention of the author's rights and establishes attribution to the authors, and the Creative Commons Public Domain Dedication license (CC-0 1.0), which allows for unrestricted public use of the work(s). For datasets, we offer users either the Open Data Commons Attribution (ODC-BY 1.0) license, which allows free share, reuse, and adaptation of data as long as authors are attributed through citations, or the Open Data Commons Public Domain Dedication License (PDDL 1.0), which allows for releasing the data for unrestricted public use. Finally, for users publishing software, including scripts, models, algorithms, etc., we offer the GNU General Public License. (GPL 3). Further detail on the license options, the data and works they can be applied to, and the rights accorded to the data authors and re-users can be found under the Licensing section in Best Practices [2]. In that same page under Subsequent Publishing are recommendations to select licenses when working on subsequent publications that require different licenses.

Awareness about licensing policy is available throughout a user's interaction with the DDR. When new users receive their DesignSafe account, they must agree to the DesignSafe Terms of Use [3], which require them to use available services in ways that will not violate intellectual property rights and licenses. Users that curate and publish data in the DDR are presented with licensing options and recommendations available in the User Guides both under Policies and Best Practices. When a user reaches the appropriate step in the publication process, they must review and select their licensing options. If users do not select a license, the publication process does not advance. Users publishing multiple and/or subsequent data products within a project can choose a different license per product (e.g., a user publishing a data report will use a Creative Commons license, and an Open Data Commons license to publish the data).

In the last step before the publication goes live, users are prompted to agree to the Data Publication Agreement [4]. Among other things, this document asks users to acknowledge the license they choose to publish their work and confirm their authority to license their work for future data transfer and preservation. They will not be able to proceed to publishing their data without agreeing to this document. The document links to the DDR Policies and Best Practices so users may fully comprehend their rights and choices. Our policy states that if we detect violations, we will notify users of any infringements they commit, and that we reserve the right to cancel the DesignSafe accounts of those that have one. See our response to R4 for more detail on user liability and consequences for violating or infringing on licenses and our user agreements, and see our response to R10 for more information on our preservation policies and the permissions required for future preservation decision making. Once the data is public, the applied licenses appear the data publication landing page.

The Data Usage Agreement establishes the DDR's expectations for those reusing data [5]. It requires users to acknowledge that they will not use data in any way that infringes on the distribution licenses selected for the data publications. Users looking to reuse data are able to review the rights carried by each publication on its landing page. In addition, the license associated with any publication appears in a pop-up screen when users download the data package from the DDR. This document also requires users to agree that they will not use data in any way that infringes on distribution licenses and permissions available for the data publication. We also include recommendations on how to reuse data in relation to its existing license in the Best Practices document under Reusing Data Sources in your Publication [6].

[1] DDR Data Licences Policy
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/policies/publication/
[2] DDR Licensing Best Practices
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/best-practices/data-publication/
[3] DesignSafe Terms of Use https://www.designsafe-ci.org/account/terms-conditions/
[4] DDR Data Publication Agreement
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/policies/publication/
[5] DDR Data Usage Agreement
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/policies/publication/
[6] Reusing Data Resources in your Publication Best Practices
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/best-practices/data-publication/

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

3. Continuity of access

R3. The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.

Compliance Level:

3 – The repository is in the implementation phase

Reviewer Entry

Reviewer 1

Comments:
3 – The repository is in the implementation phase
Accept

Reviewer 2

Comments:
3 – The repository is in the implementation phase
Accept 3.

Response:

The DesignSafe DDR has been operational since 2016 through an NSF award which was renewed in 2020 through September 30, 2025 [1]. DesignSafe is the latest awardee in the NHERI program, which NSF began in 2000 under the name NEES (Network for Earthquake Engineering Simulation). The end of the current award will represent 25 years of continuous support to the program's data. During the recent NHERI summit where the community met to discuss future directions [2], data availability and tools were identified as a top priority to advance research in natural hazards.

During the award period the DDR preserves the natural hazards research data published since its inception, while also supporting preservation of and access to legacy data and accompanying metadata from the predecessor NEES awards. The latter data collection is comprised of 33 TB, 5.1 million files, and their metadata and DOIs were transferred to DDR in 2016 as part of the conditions of the original grant solicitation (See conditions in [1]). At that time, the transition of the datasets and their corresponding DOIs was accomplished without interruption of access to the data and of curation services.

As part of the requirements of the current award, we have a continuity plan in place to transfer all the DDR data, metadata, and corresponding DOIs to a new awardee (should one be selected), again without interruption of services and access to the data. Fedora has export capabilities for transfer of data and metadata to another repository in a complete and validated fashion. This plan is a required section in the Operations Project Execution Plan that we update and present to NSF as part of our annual reporting package.

In the case in which the NSF and/or other stakeholders in this community do not continue the NHERI program or a subsequent data repository, we will continue to actively preserve the published data and will provide access to it through the Texas Advanced Computing Center (TACC), the hosting center for DesignSafe at the University of Texas at Austin. TACC has assured in writing that if project funding does not continue, the data, with landing pages and corresponding DOIs will be preserved by TACC indefinitely Should funding constraints ever make this no longer possible, we will continue to keep an archive copy on Ranch (with landing pages on online storage) for as long as TACC remains a viable entity. (See MOU attached to this application).

TACC is currently over 20 years old, and TACC and its predecessors have operated a digital data archive continuously since 1986. This archive is currently implemented in the Corral Data Management system and the Ranch tape archive system, with capacity of approximately half an exabyte. Corral and Ranch hold the data for DesignSafe and hundreds of other data collections and research projects [3]. Because TACC is constantly updating its high-performance storage

resources and security mechanisms, data will be preserved at the same preservation level that is currently available [4]. Fedora is now part of TACC's software suite, and we will continue its maintenance as our preservation repository at least for the medium term. For the long term, like with all systems at TACC, we will revisit its versioning and continuity and make decisions based on state-of-the-art practices. TACC has on permanent staff a User Services team as well as curators that will attend users' requests and help tickets related to the data. Note that all curators for the Data Depot are full time TACC employees, who dedicate part of their time to the DesignSafe contract – their jobs will continue whether DesignSafe continues or not. Considering that DOIs are supported through the University of Texas Libraries and that the web services and the data reside within TACC's managed resources, access to data will not be interrupted. This information is publicly available in our Data Preservation Policy [5].

[1] Natural Hazards Engineering Research Infrastructure (2015-2019) https://www.nsf.gov/pubs/2014/nsf14605/nsf14605.htm Note that since ince the current renewal was a no compete, and thus we are operating under the same guidelines.
[2] Natural Hazards Research Summit 2022 https://www.designsafe-ci.org/learning-center/training/nco/2022/natural-hazards-engineering-research-summit/overview/
[3] Continuing Arecibo's Legacy https://www.tacc.utexas.edu/-/continuing-arecibo-s-legacy
[4] DesignSafe Data Depot Repository Data Preservation Best Practices https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/best-practices/data-preservation/
[5] DesignSafe Data Depot Repository Data Preservation Policy https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/policies/data-preservation/

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

4. Confidentiality/Ethics

R4. The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept.

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept 4.

Response:

In the DDR we use the term protected data as an umbrella that encompasses information subject to regulations under relevant privacy and data protection laws, such as HIPAA, FERPA, and FISMA, as well as human subjects' data containing Personally Identifiable Information (PII), and data involving vulnerable populations and/or containing sensitive information. DDR's Protected Data Policy, available in [1] and the Protected Data Best Practices [2], address issues regarding confidentiality and ethics.

All data made publicly available in the DDR are intended for open reuse. At the moment, the vast majority of published data (>95%) is derived from data that have no privacy or ethical concerns with reuse. The scope of the repository has recently expanded to Social and Behavioral Sciences. While published data in the DDR is de-identified, the source data may involve kinds of Personally-Identifiable Information (PII). In those cases, data producers are required to adhere to the guidelines for long term storage of that data approved by their IRB.

Data containing PII can be published in the DDR with proper consent from the subject(s) and documentation of that consent in the project's IRB paperwork. Publishing such data involves managing direct and indirect identifiers in accordance with accepted means of data de-identification. No direct identifiers and only up to three indirect identifiers are allowed in published datasets. In the case in which removing the required direct and indirect identifiers makes data difficult to comprehend and reuse, we allow users to publish metadata about the data, and the can obtain a landing page for the publication and a DOI. The publication has to include information about how to reach the data creators in order to access the data and under which conditions. When publishing protected data or just metadata about it, we require that users include their IRB documents in the publication (under the Planning Documents category within the Field Research Data Model).

Natural hazards data may have PII in the form of granular geographical locations and images that may capture humans that are not the focus of the research and would not fall under the purview of an IRB. For example, images of research team members, people captured using a street camera during a field reconnaissance project, roofing/remodelling public records containing any form of PII, etc. For these types of data, we have recommended best practices, including blocking out or blurring any information that could be considered PII before publishing it in the DDR. Because of the diverse technologies used to capture data in the field, we coordinate with partners in the NHERI program responsible for Field Reconnaissance research (RAPID) and Social Science research (CONVERGE) to assure compliance.

In the DDR protected data issues are considered at the onset of the curation and publication workflow. Users that plan to work with human subjects should have approval from their IRB in place prior to storing, curating, and publishing data in

the DDR. At the moment of selecting a project type in the My Project interface, users are prompted to respond if they will be working with human subjects; if the answer is yes, both the DDR curator and TACC's protected data coordinator are automatically notified by email. The DDR curator gets in touch with the researchers to discuss the nature and conditions of the data and their IRB commitments. Users will also have the opportunity to review the Protected Data section of our Best Practices in the onboarding instructions, as well as the Protected Data section of our Policies in the left-hand Help drop-down menu. At the end of a project's curation and publication pipeline, users will also be prompted to agree to our Data Publication Agreement [3], which includes a statement that they are following both their IRB protocol and our Protected Data Policy and Best Practices.

As stated in the DesignSafe Terms of Use [4] and in the DDR policy, it is the users' responsibility to adhere to the requirements of their IRB and to the terms they agreed upon in their IRB presentation, as well as to the requirements they agree upon when uploading and publishing data in the DDR. Users' uploads that we verify, or are notified of, that violate these policies may be removed from the DDR, and the user may be asked to suspend their use of the DDR and other DesignSafe resources.

The DDR curators are trained in handling protected data, having taken modules assigned by UT Austin's IRB and its Office of Research Support and Compliance, and regularly refreshing their training by reviewing completed courses and taking new ones when available. In addition, TACC's protected data coordinator has undergone training programs that include relevant courses provided through CITI's IRB modules and the NIST SP 800-53 guidelines.

To manage, publish, and describe protected data in the DDR, we developed the interdisciplinary Field Research data model in coordination with our CONVERGE partners [5]. This model contains metadata elements that map to the Data Documentation Initiative (DDI) metadata standard including the description of access restrictions [6].

Privacy and ethical concerns are evolving issues, and we are continuously monitoring and discussing how to address data gathering and anonymization with the CONVERGE team, among other things to avoid opportunity for deductive disclosure of people inadvertently captured in field research data. For training purposes, CONVERGE has a series of check sheets [7] that outline how researchers should ethically manage data that contain personal and sensitive information; these check sheets have also been published in the DDR.

We request that users reusing data do not obtain personal information associated with DDR data that results in directly or indirectly identifying research subjects, individuals, or organizations with the aid of other information acquired elsewhere. This is stated in the Data Usage Agreement [8] that users agree to as they download and reuse the data.

[1]DDR Curation and Publication Policies, Protected Data
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/policies/publication/
[2] DDR Curation and Publication Best Practices, Protected Data
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/best-practices/data-publication/ .
[3] Data Publication Agreement
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/policies/publication/

[4] DesignSafe Terms of Use https://www.designsafe-ci.org/account/terms-conditions/
[5] CONVERGE Interdisciplinary field research data model https://converge.colorado.edu/data/data-publication
[6] Data Documentation Initiative. https://ddialliance.org/products/overview-of-current-products.
[7] CONVERGE Check Sheets https://converge.colorado.edu/resources/check-sheets/ethical-considerations/
[8] Data Publication and Data Usage Agreements
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/policies/publication/

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

5. Organizational infrastructure

R5. The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept.

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept 4

Response:

This data repository has a history of community continuity [1]. DesignSafe is the fourth in a series of awards to steward NSF’s natural hazards engineering data over a span of 25 years. Much of the DesignSafe activity resides at the Texas Advanced Computing Center (TACC), which is in turn an Organized Research Unit within the University of Texas at Austin.

DesignSafe was originally funded in 2015 via a five-year award from the National Science Foundation (NSF), and was renewed beginning October 1, 2020 for five additional years. The DDR has been allocated 2.5 FTE and $1.96M of the personnel budget of the NSF award. Separately, there are funding lines for travel and workshop/conference registrations that are associated with our staff both providing and receiving training and professional development. The budget also includes infrastructure funding for the disk storage, and virtual machines (VMs) hosted at TACC that underpin the DDR. This budgeting is reviewed and necessary adjustments are made annually within the context of an NSF review.

DesignSafe is one project within the larger TACC organization [2]. TACC is more than 20 years old, with more than 180 full-time staff and annual expenditures of >$40M/year. UT-Austin is nearly 140 years old, and has an annual operating budget in excess of $3B. The staff responsible for the DDR reside in groups of like-minded professionals at TACC, spread across the data group for curation and data management (a total of ~12 FTE), the Interfaces group for front-end software and UI design (approximately 35 FTE), and the Systems group for running actual server and storage systems (~25 FTE). This assures sufficient redundancy and continuity beyond the current staff assigned to DesignSafe, and ensures that these staff work in groups of similar professionals working on a range of related projects, providing broad exposure to the best practices in the field. All staff working on the repository are UT/TACC employees, with positions that exist independent of the funding for DesignSafe.

Below is a breakdown of the 2.5 FTE allocated to the DDR across the areas of expertise the DDR staff constitutes, and more information on how these areas of expertise inform DDR operations can be found in R6.

Expertise FTE Level

Data Curation 0.5
Software Development 1.0
Web Design 0.15
Training and User Support 0.2
Usability/User Experience 0.6
Administrative Management 0.05

TOTAL 2.5

The team working on the DDR [3] operates under the leadership of the DesignSafe Management Team [4], who establishes and updates requirements and policies, recommends and tests best practices, oversees technical developments, and prioritizes activities. To operationalize activities, both the DesignSafe PI and its Deputy Director have formal monthly meetings with the other awardees in the NHERI community, including NSF-funded experimental facilities, and field reconnaissance teams who publish data through the DDR. The management team meets on a bi-weekly basis. Issues regarding the DDR are communicated across these channels to and from the repository team, recorded, and

transformed into action items by DesignSafe's project manager.

The Repository Team [3] represents backgrounds in information and computer sciences, web development, systems administration, natural hazards engineering, and user experience fields, gathers requirements, proposes and designs solutions, and carries out ongoing design, development, and day-to-day operations.

The lead data curator has a PhD in Information Sciences with focus on digital preservation. She has worked at TACC since 2008, developing data repository platforms and large data curation workflows in connection to High Performance Computing resources. In addition to participating in the NHERI and DesignSafe Management Team meetings as needed, the data curator meets twice a month with the DesignSafe PI and the lead portal developer.

The data curator establishes the technical standards for the repository and conducts regular work meetings with the rest of the team to communicate those and to evaluate development. In collaboration with the User Experience staff member, they provide all the mock ups, instructions, vocabularies, and documentation required for the interface. They also review data publications monthly to learn how users are using the interactive pipelines and onboarding instructions. The curator serves as a bridge between domain scientists, the analysis activities performed on DesignSafe, and the DDR to assure seamless experiences in data. Both curators and usability experts conduct virtual office hours where they assist users and receive feedback and conduct training through webinars. Different members of the development team respond to help tickets and assist users with technical problems through Slack. They meet weekly with the lead portal developer to work through the agenda of curation and publication tasks which are organized in 90-day cycles. All team members attend ad-hoc meetings and maintain regular communications with the broader members of the network via Slack or email. All tasks and their characteristics and progress are recorded in a Jira instance.

Team members have the opportunity to attend training as needed, and present their work in professional meetings as well as publish in collaboration with the domain scientists. Below are a sample of latest publications and presentations by the repository team:

** Jean Paul Pinelli, Maria Esteva, David Roueche, Jamie Padgett, Gilberto Mosqueda, Frederick Haan, Scott Brandenberg, Ellen Rathje (2020). Disaster Risk Management through the DesignSafe Cyberinfrastructure. International Journal of Disaster Risk Science. DOI: https://doi.org/10.1007/s13753-020-00320-8.
** Maria Esteva, Ellen Rathje. Perspectives from Data Reuse in the Field of Natural Hazards Engineering. 12th Qualitative and Quantitative Methods in Libraries International Conference (QQML 2020) 26-30 May, Barcelona, Spain. http://qqml.org/wp-content/uploads/2017/09/Book-of-Abstracts-26-5-2020-.pdf
** Rosenberg, Jake, et al. "Leveraging Elasticsearch to Improve Data Discoverability in Science Gateways." Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (Learning), Association for Computing Machinery, 2019, pp. 1–5. ACM Digital Library, doi:10.1145/3332186.3332230.
** Keith Strmiska, et al. (2019) Simplifying Natural Hazards Engineering Research Data with an Interactive Curation Process. Science Gateways 2019, San Diego, CA, USA, September 23-25.
** Maria Esteva, Craig Jansen, Josue Balandrano Coronel (2019). Designing and Building Interactive Curation Pipelines for Natural Hazards Engineering Data. International Journal of Digital Curation. Vol 13, No.1, DOI: 10.2218/ijdc.v13i1.661

[1] DDR Mission and History https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/policies/

[2] TACC https://www.tacc.utexas.edu

[3] DDR Governance https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/policies/governance/

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

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

6. Expert guidance

R6. The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either inhouse or external, including scientific guidance, if relevant).

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept.

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept 4.

Response:

In the DDR we consider expert guidance from internal and external advisors. Any advisor that is funded under the DesignSafe NSF award is considered internal, and any other advisor that is a member of the broader NHERI community, is contracted to provide assessments and evaluation, or is part of the DesignSafe Advisory Committee, is considered external.

Internal advisory meetings in relation to the DDR include the domain and technical experts in the DesignSafe Management Team, as well as the members of the Use Case, Implementation, and Community Engagement Teams. See the DesignSafe Organization at [1]. During the first five years of DesignSafe operations, the repository and the internal advisors conducted bi-weekly meetings as a Data Management cluster which was instrumental in the design and evaluation of the DDR data models and in defining the policies that govern the repository. Currently, meetings are called every quarter to discuss evolving issues and new themes such as requirements for machine learning and resilience data, the data awards, treatment of PII information in large scale geoinformatics data, etc.

The repository maintains regular communications with members of the NHERI network described in R0. This contact includes bi-monthly meetings between the DesignSafe team and the rest of the NHERI network awardees. There is a strong culture of collaboration between DDR and the NHERI team. During the first years of DesignSafe, the DDR staff visited all the EFs and the SimCenter, attended field data training workshops from the RAPID, and observed field research work. Most recently, the outcomes of these community engagement included the release of onboarding data curation instructions for experimental facilities [2] and publish your data events [3]. Data news are published by NHERI on a regular basis to keep the community informed and engaged.

Virtual office hours are a fundamental venue for the DDR curators to gather feedback from the community. Many questions and suggestions concerning the DDR are later implemented in the curation and publication workflows including: adding controlled terms to the tagging vocabularies, improving guides, adding instructions, and testing/fixing new releases. The DDR team also consults with resources and staff from UT Austin as needed, primarily from UT Information Technology Services (ITS), its Office of Research Support and Compliance (RSC), and the UT Libraries.

Formal mechanisms and funding are in place for external evaluators to gather feedback and conduct structured assessments, in the form of usability studies and yearly user surveys, to ensure that the repository is meeting the community’s expectations and needs. For example, previous to their release, the interactive curation and publication pipelines have been iteratively evaluated by user experience consultants with the participation of users from the community during a professional conference.

DDR activities are reported to the National Science Foundation on a quarterly and annual basis in terms of quantitative and qualitative progress. The annual review consists of a site visit during which the progress of the team is presented, and the review panel conformed by experts in all the areas relevant to DesignSafe makes recommendations which are acted upon during the year.

DesignSafe has recently applied for membership in the EarthCube Council of Data Facilities [4], a forum for advancement of data in the earth system science domains.

[1] DesignSafe organization https://www.designsafe-ci.org/about/designsafe/.
[2] DesignSafe Experimental Facility Onboarding Checklist for Data Curation https://www.designsafe-ci.org/rw/user-guides/managing-data/ef-checklist/.

[3] CONVERGE-DesignSafe Publish Your Data Workshop https://converge.colorado.edu/data/events/publish-your-data.
[4] Earthcube Council of Data Facilities https://www.earthcube.org/council-of-data-facilities

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

DIGITAL OBJECT MANAGEMENT

7. Data integrity and authenticity

R7. The repository guarantees the integrity and authenticity of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept.

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept 4.

Response:

The DDR is implemented within the DesignSafe cyberinfrastructure which provides end-to-end cloud-based tools and secure storage to manage, understand, and publish data for critical natural hazards research. In this context, the DDR functions as a space that connects data to tools and applications for analysis, as well as to curation, publication, and archiving services.

The data in the repository is maintained in the broader DesignSafe workspace with a separate preservation back end to ensure its integrity. The preservation backend of the DDR is a Fedora repository [1]. The Fedora repository is a digital library standard that complies with the OAIS Reference Model [2]. Fedora has functionalities to assure the integrity and

authenticity of the digital objects stored within. DDR data are stored in a Duraspace Fedora 5.x. Repository as binary assets and RDF metadata.

The integrity and authenticity of the data in the repository is achieved through processes - seamless to the users - that spans the DesignSafe front-end and the preservation back-end. To upload and manage data in DesignSafe, users need to obtain an account, and to share data within an active project the project principal investigator has to authorize registered team members. User credentials are validated before performing any operation.

Using interactive curation functionalities in the DDR web interface, users upload, organize, describe and tag their datasets. For publishing data, users select files, review their descriptions and content, select a license, read and approve the publication agreement, and publish their data (see the interactive process in the Curation and Publication User Guide [3]). Across relevant steps, the system conducts automated verification of completeness of the data and metadata, and users are not able to proceed if critical metadata or data are missing. Once the user has concluded and reviewed the publication steps, they can request publication. At this step, several actions are taken:

• A read-only copy of the user's data is made in the web front end. This is the visible online copy – our experience for large datasets (our repository has millions of files and tens of terabytes of data) is that data stored directly in Fedora is not sufficiently performant for our user needs.

• A copy of the publication data is compressed and used for download. This Dissemination Information Package (DIP) includes a metadata file.

• From the read-only copy, data and metadata about the data publication are submitted to Fedora as a Submission Information Package (SIP) behind the scenes.

Fedora is the repository for the archival information package (AIP). The Fedora Digital Object Model [4] design assures maintenance of provenance through relations between the data streams, metadata, identifiers, and the eventual versions. The Fedora instance is not directly accessible to users, only to administrators. At ingest, Fedora extracts file format information and conducts transmission fixity checking [5]. In our implementation, persistence fixity checking is performed once a year in staggered fashion. Per data object, file format is stored as ebucore metadata and fixity information as PREMIS Message Digest property [6]. Once the data has been published, the user may not change the data in the repository again. While new versions and amendments may be published, the original package is preserved as of the date of publication, and so are the subsequent versions [7].

[1] The Fedora repository https://duraspace.org/fedora/about/.
[2] OAIS Reference Model http://www.oais.info/oais-usage/.
[3] DesignSafe Curation and Publication User Guide
https://www.designsafe-ci.org/rw/user-guides/data-curation-publication/.
[4] Fedora Digital Object Model

https://wiki.lyrasis.org/display/FEDORA34/Fedora+Digital+Object+Model#FedoraDigitalObjectModel-Datastreamsdata

[5] Fedora Fixity Checking https://wiki.lyrasis.org/display/FEDORA51/Fixity+Checking

[6] Fedora Metadata Recommendations https://wiki.lyrasis.org/display/FEDORA5x/Metadata+Recommendations

[7] Fedora Versioning https://wiki.lyrasis.org/display/FEDORA51/Versioning

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

8. Appraisal

R8. The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept.

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept 4.

Response:

The DDR has published Collection Development policies in relation to Data Types, Data Sizes and Data Formats [1]. We accept engineering and social and behavioural sciences datasets (SBE), derived from research conducted in the context of natural hazards. In the area of engineering the primary focus is on data generated through simulation, hybrid simulation, experimental and field research methods regarding the impacts of wind, earthquake, and storm surge hazards. We also accept data reports, publications of Jupyter notebooks, code, scripts, lectures, and learning materials. In SBE, accepted datasets encompass the study of the human dimensions of hazards and disasters. As the field and the expertise of the community evolves, we have expanded our focus to include datasets related to COVID-19, Fire Hazards, and Sustainable Material Management.

Since the DDR's inception there have not been any cases where data publications have fallen outside of the accepted collection profiles. To guide users, we provide them with our Mission Statement, Data Collection Development Policy, and Curation and Publication User Guide so they can determine whether their work would fit in the DDR. While the DDR is a self-archiving repository, our curators review in-process and published datasets regularly as detailed in R11. Our policy states that users that store data which does not correspond to the accepted types will be alerted. If possible, this happens prior to publication. If a dataset that is non-compliant with the Collections Development policy gets published, we will alert the authors and leave a tombstone [2]. In both cases we will work with the authors to find an adequate repository for the dataset.

To date, the only metadata schema in the field of Natural Hazards Engineering are those developed within DesignSafe. Beginning with the previous NEES data repository the community has gathered requirements, committed to data best practices, and achieved experience documenting the complex research projects that they conduct [3]. Building on these experiences, during the first five years of the DDR operations we worked closely with the domain scientists to develop data models that specify the data, the documentation, and the metadata terms (DDR metadata) for data publications derived from simulations, experiments, field research and hybrid simulations performed by natural hazards engineers [4, 5]. For social science datasets, we were able to leverage existing work and developed a metadata profile based on the Data Documentation Initiative (DDI). The terms selected and agreed upon by the domain scientists to describe the datasets come from the technical vocabulary that the natural hazards community is familiar with.
For purposes of metadata exchange, to ingest Submission Information Packages (SIP) in Fedora, to use as Search Engine Optimization metadata, and to mint DOIs, we have mapped the DDR data models' terms (metadata) to Dublin Core, DataCite, DDI and PROV schemas. In turn, Fedora natively captures preservation metadata in PREMIS, which assures that we can retrieve relevant representation information for long-term preservation.

The Data Model and Metadata Policies are published at [6], and the Metadata Requirements including a dictionary and how the terms map to standard metadata schemas for all data models are available at [7].

We facilitate comprehension of and compliance with metadata requirements both for users publishing data and for those using the published data. For the former, we provide the DDR metadata definitions and descriptions of the types of files required for data completeness within the curation and publication pipelines onboarding instructions. We also provide guidance of how to write descriptions and highlight which are required fields. To facilitate adherence with metadata completeness, the publication pipeline conducts automatic checks and alerts users when required metadata is not present. We also require that users include a data report to clarify aspects of data organization and description. See the Curation and Publication User Guide to follow the curation and publication processes at [8]. Based on the data models we designed landing pages to help users navigate and understand the structure and content of the data publications. Our design approach was for users to visualize the correspondence between data and the processes from which it derives [9]. The landing pages were tested for comprehension/usability by real users and by an independent user experience consultant and modified accordingly.

As users publish more datasets, we continue assessing metadata and publication understandibility, consulting with stakeholders, and building upon the data models and their corresponding vocabularies. We are in communication with our community about these practices through virtual office hours, training webinars and best practices documentation.

As explained in our Collections Development policy, in attention to evolving advances in methods and instrumentation as well as to established research data practices in our community, we do not have strict file format requirements for the engineering and social science datasets uploaded to the DDR. However, we encourage and educate our users to upload their data following the Library of Congress Recommended Formats for long-term preservation and interoperability purposes. Our Accepted and Recommended File Formats Best Practices [10] and our General Research Data Best Practices [11] highlight our preferred formats, including links to data curation primers. Many of our users generate and publish complex dimensional models, scripts and algorithms, as well as analytic results. Some of these data types do not translate well when converted to open formats, and in particular the engineering community considers software such as Matlab and ArcGIS as standards in their educational and professional practices. To facilitate both reuse by the community and preservation, we allow users to publish their datasets both in proprietary and preservation friendly formats. As part of the publication process, file formats are identified in the Fedora repository which generates PREMIS metadata for each file, allowing us to monitor file formats in our repository.

We do not pose file size/project size limits. It is not unusual for publications in the Natural Hazards space to contain hundreds to many thousands of files, and scale to the terabytes. The large size and the complexity of the projects led us to create interactive data curation interfaces that would facilitate organizing and representing these projects. See our Data Size Policy [12] and our Data Size Best Practices [13]. In addition, we do provide recommendations about data selection in regards to size and reuse considerations in our Data Quality Control Best Practices [14].

[1] DDR Collection Development Policies
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/policies/collections/
[2] DataCite Tombstone Best Practices https://support.datacite.org/docs/tombstone-pages.
[3] Building Infrastructure for Preservation and Publication of Earthquake Engineering Research Data
https://doi.org/10.2218/ijdc.v9i2.335.
[4] Yield Surface Mapping and Triaxial Compression Test Data Curation. https://doi.org/10.1061/9780784481585.022
[5] Curation and Publication of Simulation Data in DesignSafe, a Natural Hazards Engineering Open Platform and Repository. https://doi.org/10.3390/publications7030051
[6] DDR Data Models and Metadata Policy
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/policies/curation/.
[7] DDR Metadata Requirements Best Practices
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/best-practices/data-curation/.
[8] DDR Curation and Publication User Guide https://www.designsafe-ci.org/rw/user-guides/data-curation-publication/
[9] Designing and Building Interactive Curation Pipelines for Natural Hazards Engineering Data. DOI
10.2218/ijdc.v13i1.661

[10] DDR Data Collections Development Best Practices
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/best-practices/
[11] DDR General Research Data Best Practices
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/best-practices/data-curation/
[12] DDR Data Size Policy https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/policies/collections/
[13] DDR Data Size Best Practices
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/best-practices/
[14] DDR Data Quality Control Best Practices
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/best-practices/data-curation/

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

9. Documented storage procedures

R9. The repository applies documented processes and procedures in managing archival storage of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept 4.

Response:

The underlying storage systems for the DDR are managed in-house at the Texas Advanced Computing Center (TACC).

All the storage systems used by DesignSafe are shared multi-tenant systems, hosting many projects concurrently in

addition to DesignSafe – the front-end disk system currently has ~20PB of data, with the tape archive containing roughly 80PB [1, 2]. These systems are operated in production by a large team of professional staff, in conjunction with TACC's supercomputing platforms. Public user guides [3,4] document the capabilities and hardware. A Redmine site, visible only to systems staff, is used to document and maintain procedures for staff, including configuration management of the scripts that automatically ingest user data into Fedora. In the case of the DDR, filesystem replication is automatic. Ingestion of data from the web-visible storage into Fedora takes place under automated control when the publication workflow executes. The Fedora repository and database are likewise replicated as well as backed up on an automated schedule.

All storage locations are managed in house; we know exactly which rack in which data center maintains each copy of the data. Both the front-end copies and the Fedora repositories are in systems that implement de-clustered RAID and have sufficient redundancy to manage up to 3 drive failures for a single file stripe. The file system itself is mirrored daily between two data centers. The primary data is also periodically backed up to a tape archive for a third copy, in a third data center. The database that manages metadata in Fedora is also quiesced, snapshotted, and backed to tape on a regular automated schedule.

The primary copy of the published data is made for ingest into Fedora which generates checksums on each file at ingest. On a rotating schedule, but at least once per year, new checksums are generated from the online copy to guarantee that these still match.

Corruption in the disk system would be detected immediately due to the filesystem mirroring. All data on tape media are periodically re-read and written to newer tapes – roughly every 3rd generation of LTO tape release (typical drives can read the current generation and two previous ones – we retire all of the oldest generation tapes before replacing drives).

Risk management is an integral part of the overall project and center operations. Our risk registers around operation of the storage systems are reviewed in TACC's annual operations reviews with the National Science Foundation.

[1] Corral https://www.tacc.utexas.edu/systems/corral
[2] Ranch https://www.tacc.utexas.edu/systems/ranch
[3] Corral User Guide https://portal.tacc.utexas.edu/user-guides/corral
[4] Ranch User Guide https://portal.tacc.utexas.edu/user-guides/ranch

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

10. Preservation plan

R10. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept.

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept 4.
Certain for the duration of the NSF project (up to 2025) and strengthened by the MOU between TACC and Designsafe after 2025.

Response:

Our Data Preservation Policy is publicly available at [1], and the Data Preservation Best Practices can be found at [2]. These documents are intended to convey information to the depositor. The preservation approach internally is defined in our internal document management systems, maintained in Confluence.

Data in the DDR are maintained according to state-of-the art standards and best practices in what we define as a preservation environment. DesignSafe is implemented within the reliable, secure, and scalable storage infrastructure at the Texas Advanced Computing Center (TACC), with 20 years of experience and innovation in High Performance Computing. Within TACC's storage infrastructure, a Fedora repository - a standard in the Digital Library community that complies with the OAIS model - manages the archival information package (AIP) and enables preservation of the published data. Within Fedora, the authenticity and integrity of each digital object is accounted for and tracked, as well as its relations with the rest of the data and metadata that encompasses a publication. This preservation environment allows us to maintain data in secure conditions at all times, before and after publication, comply with, NDSA Preservation Level 1 [3], attain and maintain the required representation and descriptive information about each file, and be ready at any time to transfer the custody of published data and metadata in an orderly and complete fashion. Fedora has export capabilities for transfer of data and metadata to another Fedora repository or to another system [4]. The underlying storage infrastructure is updated on a roughly 6-year cadence for online data. For data stored on tape, we maintain tape drives for 3 generations of media, and copy all data to new media before expiring a particular type of tape (typically 5-8 years).

Depositors of data in DDR go through a curation and publication pipeline that sets – for each type of data model – the standard of metadata and content requirements for data submission and automatically checks that those requirements are

met (See guides for curation and publication in [5]). In turn, the pipeline contains links to the published policies and documentation that support those requirements.

Users that deposit and publish datasets in DDR are informed about our commitment to preservation and access when they acknowledge and accept the Data Publication Agreement [6] at the end of the data publication pipeline. In it, we clearly ask the users to grant the licenses and permissions to make the data available and to “store, translate, copy, transfer or reformat files in any way to ensure its future preservation and accessibility” .
Depositors are aware of the requirements for publication -submission information package- through a well- documented onboarding curation pipeline whose functionalities including automated checking that metadata and content are complete, are explained in the Curation and Publication User Guide [4]. In turn, the data preservation process by which objects and metadata are ingested to Fedora which maintains the archival information package are explained – for the public – in the Data Preservation Best Practices [2].

The DDR Fedora repository captures, maintains and updates the preservation metadata of the published archival datasets (archival information packages in OAIS terms) through PREMIS. This information is relevant to any future preservation action including transfer to a future awardee.

In the Data Preservation Policy [1], we inform depositors about the long-term preservation of the datasets in compliance with the requirement posted by the National Science Foundation (NSF) in the original Requests for Proposal that states "If transition to a different platform becomes necessary in the future, then the CI Awardee will be responsible for ensuring that all content, software, and tools are fully transitioned to that platform without requiring renegotiation of proprietary agreements." In addition, in the last paragraph of the policy we reassure them that their data will be preserved in the unlikely case that the funded program is discontinued.

Our cooperative agreement with the NSF [7] specifies our responsibility and commitment to transfer the data and metadata to the next awardee. Our process is to use relevant Fedora tools to assure the transparency and integrity of the collection. This process will also include transferring all the DOIs according to the procedures established by DataCite [8]. We have experience in this process having gone through it in 2016 when data, metadata, and corresponding DOIs were transferred from the Network for Earthquake Engineers Simulation (NEES), the former iteration of this award, to DesignSafe. Data and metadata transfer to another awardee after the grant period is a requirement of the grant proposal

[1]Data Preservation Policy
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/policies/data-preservation/
[2]Data Preservation Best Practices
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/best-practices/data-preservation/
[3] Levels of Digital Preservation. National Digital Stewardship Alliance – Digital Library Federation, http://ndsa.org//publications/levels-of-digital-preservation/.
[4] Fedora import and export tools https://github.com/fcrepo-exts/fcrepo-import-export
[5] Curation and Publication User Guides https://www.designsafe-ci.org/rw/user-guides/data-curation-publication/[6] DDR Data Publication Agreement

https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/policies/publication/.
[7] National Science Foundation Award Abstract # 2022469 Natural Hazards Engineering Research Infrastructure: Cyberinfrastructure (DesignSafe) 2020-2025 https://www.nsf.gov/awardsearch/showAward?AWD_ID=2022469
[8] DataCite Transfer DOIs Documentation https://support.datacite.org/docs/fabrica-transfer-dois.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

11. Data quality

R11. The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept.

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept 4.

Response:

The NHERI program is governed by a council run by the coordinating office, consisting of the various awardees that make up the program: the Experimental Facilities, CONVERGE, RAPID, the SimCenter, and DesignSafe [1]. The decision of the council was that end users (e.g. the data producers) would be responsible for the quality of their own results, and that our repository would look only at the completeness of the data, and the quality and completeness of metadata for publication. The community standard was to provide a “data report” file to discuss the methodology used in the experiments themselves, beyond what is normally captured in the metadata as well as the recommendation to append the relevant papers published about the research that originated the data. Metadata and Data Quality Policies and Best Practices

outlining requirements and providing guidance are publicly available [2, 3]

Only registered users, approved by vetted project PIs, may upload data or metadata for publication.
Metadata is entered via one of five schemas developed with the community (experimental, simulation, hybrid simulation, field research, and other). The DDR offers suggestions on how to fill metadata entries within the metadata entry forms, controlled vocabularies to fill some of the information in these fields, as well as onboarding guidance throughout the curation and publication pipelines that explain the meaning of metadata fields and suggest which data types and documentation files should be uploaded [4]. Curators spot check accuracy of the content via regular reviews and office hours, and they contact users as needed for metadata clarification and enrichment.

Different groups in the NHERI network have produced extensive documentation about data curation and standards of practice in relation to data validation. We work closely with them both to develop and share these documents [5, 6, 7].

The automated checks within the system ensure that there is information in relevant metadata fields that block progress towards publication if they are not completed. In addition, if data files are missing from key categories, signalling that the project is incomplete, the system will not proceed to publication. In both cases, appropriate error messages are generated and shown to the users for correction. Curators review newly published projects, and those submitted to be published, to examine metadata and documentation. Online conversations and/or help ticket exchanges with data depositors are used to resolve any issues prior to issuance of the DOI. In the event corrections are needed to published data, we have a facility for amendments or publication of new versions [8, 9].

Users publishing data can include any citations they wish as part of the metadata of their dataset within Related Works and Referenced Data fields. We use the facilities of DataCite to relate DOIs of related publications within each publication’s metadata. We include extensive documentation to stimulate users to include references and better contextualize their publications. [10, 11, 12]

Users can click a “Leave Feedback” button on the projects’ landing pages to provide comments on any data publication. This feedback is forwarded to the curation team for any needed actions, including contacting the authors. In addition, it is possible for users to message the authors directly as their contact information is available via the authors field in the landing pages. We encourage users to provide positive feedback and suggest themes they may want to discuss about the publication in our Data Feedback Best Practices [13].

More broadly, we have regular meetings (monthly for the first five years, now quarterly) with internal and external advisors in the community regarding curation, metadata schemas and their implementation/evaluation, data quality, and suggested revisions to our policies and best practices. An output of these meetings includes the EF onboarding check-list which assures that a meeting with the curator takes place during curation and prior to publishing data [14].

[1] NHERI Network https://www.designsafe-ci.org/about/
[2] Metadata and Data Quality Policies
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/policies/curation/

[3] Project Documentation and Data Quality Best Practices
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/best-practices/data-curation/
[4] Metadata Requirements
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/best-practices/data-curation/
[5] Guidance for Data Management Plans, Converge Extreme Events Research Check Sheets Series
https://doi.org/10.17603/ds2-ycz2-xc47. Also available for download at:
https://converge.colorado.edu/resources/check-sheets/ See Guidance for Data Management Plans.
[6] STEER Product Curation Handbook https://drive.google.com/file/d/15EO6WfrUlq8vnrKMebYJQgTWEIw7aNFc/view
[7] RAPID Data Publishing Guidelines https://rapid.designsafe-ci.org/resources/
[8] Amends and Version Control Policies
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/policies/publication/
[9] Amends and Version Control Best Practices
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/best-practices/data-publication/
[10] FAQs Data Reuse https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/faq/
[11] Data Citation Policies https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/policies/publication/
[12] Reusing Data Resources in your Publication Best Practices
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/best-practices/data-publication/
[13] Leave Data Feedback Best Practices
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/best-practices/data-publication/
[14] Experimental Facility Checklist https://www.designsafe-ci.org/rw/user-guides/managing-data/ef-checklist/

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

12. Workflows

R12. Archiving takes place according to defined workflows from ingest to dissemination.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept.

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept 4.

Response:

Workflows for the DDR are thoroughly documented across several user guides and in our Policies and Best Practices. From the portal's main webpage through Getting Started, or navigating directly to the User Guides [1] users will find all the information that they need to begin using the DDR; from uploading data using different data transfer methods, to curating and publishing the datasets in their projects, to finding and retrieving data and metadata in the DDR. All of these guides can be found under the 'Help' tab on the DesignSafe menu across the top of the webpage.

We can trace how our workflows correspond to the Open Archival Information System (OAIS) reference model [2].

Submission Information Package (SIP)

Users create the SIP through the processes of uploading data to My Projects, selecting a project type, and completing the process of curation and publication.

Organizing and describing the different project types available in the DDR (experimental, simulation, hybrid simulation, field research, other) follows a similar pathway and consistent user interface design, albeit different metadata schemas and provisions for protected data. We describe how we advise users on managing protected data and how those protocols are integrated into the workflow of the DDR in R4. The curation steps for each project type can be followed in the Curation and Publication Guide [3]. Data is managed and curated in My Projects where there are three spaces with different and related functions. In the Working Directory users upload, copy, organize, review all the files/data they have deposited into My Project. The Curation Directory tab allows data organization and description according to the different data models and controlled vocabularies. The Publication Preview tab allows users to review the curation results and proceed to data publication, amends and versioning.

Throughout the workflow users will find guidelines and onboarding instructions at every step. When users create their project in the "Edit Project" menu, each field features a description of what information the user should provide and whether the field is required; users will not be allowed to proceed to the following step without responding to every required field on the current page, and they can always go back and change or add their responses. The "View Overview" page located next to the project type in the project menu allows users to review a description of the project type they selected and curation guidance, including step descriptions, definitions of metadata terms, and links to the Curation & Publication Guide, Best Practices, Help ticketing system, and the DDR curator's office hours. Similar guidance is available as users describe the experiments, simulations, field missions and hybrid simulation components of their projects. The final step to complete the SIP is for users to agree to the Data Publication Agreement [4]. After this step, the workflow to

mint the DOI in DataCite runs automatically and a DOI is assigned. Data and metadata – including the information about the DOI constitute the SIP that is submitted to Fedora through its REST API in automated steps. If users produce amends or versions, metadata and files are submitted to Fedora automatically.

Archival Information Package (AIP)

When data and metadata are ingested into the Fedora repository, user-provided metadata is mapped to the standard schemas available in Fedora (ebucore, PREMIS, DublinCore and PROV). Fedora identifies the file formats, generates checksums for each file, generates and stores metadata in RDF. The checksums are recomputed and compared to the original copy at least annually (see R7).

Fedora manages the AIP through audit trail, version control, and PREMIS event metadata updates.

The curation team performs a monthly review of new published datasets to add a human check for errors as well as to evaluate compliance with best practices and user experience. If issues are detected by curators or by data producers, users can amend and or version their publication and each new version, including amends, are tracked and preserved in Fedora.

Distribution Information Package (DIP)

As we mentioned in R7, a read-only copy of the published data is made available for distribution (viewing and download) in the DDR, as data stored in Fedora is not performant for our needs due to the large size of the datasets generated in this community. Each data package available for download includes a metadata file in JSON format. The AIPs in Fedora (data and metadata) will be used as DIPs in the case of data transfer to another repository or institution. The metadata generated by Fedora for each publication will be exchanged in the DDR and available for download with the zipped data packages in the projects landing pages. More information on how Fedora manages digital objects is available in its documentation [5].

Over time we have changed our interactive interface design to incorporate new features and terms, or to change those that did not work in response to community experts, feedback from users, and usability testing. This iterative process will continue as we welcome new communities that have different requirements, and as we learn what users need from our studies and formal and informal user feedback. Understanding that changes have an impact on the users understandibility of the platform and the workflows, new releases are explained to users by updating the onboarding instructions, the User Guides and when applicable, communicated through general emails to the user community. New features are continuously added to the back-end to strengthen the services and infrastructure, although these are not obvious to the users.

All changes done to the workflows in the interface and to the back-end are documented internally for the DDR team in a DesignSafe JIRA instance. The overall process and procedures for this workflow are documented in our internal Confluence site, to insure that staff have access to the full workflow and to the intent behind each step as staff inevitably turnover. This information is password-protected and available only to the implementation team, though the general overview of the process is available in the user-level documentation. Regular meeting notes are also kept here to record

discussions of updates, and formal configuration changes to any part of the workflow or the underlying software components are captured in our JIRA site.

[1] DesignSafe User Guides https://www.designsafe-ci.org/rw/user-guides/
[2] OAIS Reference Model http://www.oais.info/
[3]Curation and Publication Guide https://www.designsafe-ci.org/rw/user-guides/data-curation-publication/
[4] DDR Data Publication Agreement
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/policies/publication/
[5] Fedora Documentation https://duraspace.org/fedora/resources/

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

13. Data discovery and identification

R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept.

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept 4.

Response:

The DDR provides users with search functionalities. Users can search through all publications in the public data browser (https://www.designsafe-ci.org/data/browser/public/) by author, title, keyword, full text descriptions, as well as by other

metadata of relevance to this community such as project type. To refine their search users can apply different filters such as natural hazard type and event name, experimental facility, types of experiments, simulations and hybrid simulations, and data types. The search engine is built and managed through Elastic Search using the metadata that the users contribute through the web interface in relation to our data models.

As explained in our Data Model and Metadata Policies [1], to facilitate data curation of the diverse and large datasets generated in the fields associated with natural hazards, we developed five data models and metadata for experimental, simulation, field research, hybrid simulation, and other data products. We also gathered lists of controlled terms applicable to each data model. Based on the Core Scientific Metadata Model, [2] these data models and metadata were designed in collaboration with our expert community and considering their research practices and workflows, their needs for documenting research data processes, and using discipline-specific terms [3, 4, 5, 6]. The models/metadata highlight the structure and components of natural hazards research projects across time, methods, geographical locations, provenance, and instrumentation. In turn, the model's metadata is mapped to standard schemas used in the field (DDI, DublinCore, DataCite and PROV). The metadata mapping for each model is documented in [7]. This mapping is used for ingesting SIPS (data and metadata) to Fedora.

The repository provides Digital Object Identifiers (DOIs) issued by DataCite – which relies on the DataCite metadata schema for accurate citation - for all published datasets. This service is provided through UT Libraries which manages and funds the contract with DataCite for all UT Austin. DesignSafe has one of the accounts and implements the pipeline to mint DOIs with the unique DS2 suffix. Our Data Citation Policy [8] and the Curation and Publication FAQ [9] provide citation guidance to users publishing data projects that include multiple citable components in the DDR. We also enable referencing related works and/or data reused in their projects through the Related Work and Referenced Data fields. Through our Data Usage Agreement [10], available when downloading data, we communicate to users that proper attribution is one of their responsibilities when using data from the DDR.

We encourage and facilitate researchers using data from the DDR to cite it using the DOI and citation language available in the datasets landing page- which they can copy paste and download in DataCite format. While we cannot know with certainty if every user complies, our approach is to educate our community by reinforcing citation in a positive way. For this we implement outreach strategies to stimulate data citation; through the FAQs, webinars, and via emails, we regularly train our users on data citation best practices. And, by tracking and publishing information about the impact and science contributions [11] of the works they publish citing the data that they use, we demonstrate the value of data reuse and further stimulate publishing and citing data using persistent identifiers.

We have done extensive search engine optimization (SEO) work to make our data more discoverable by Google and other search engines, and include many metadata tags embedded in the HTML of the data publication landing pages. As a result, our datasets are findable through search engines shortly after publication, and are quickly indexed by Google Data Search both through DataCite and due to our SEO work. We explain our Data Impact Policy [12] and provide best practices recommendations for marketing data [13] in our public documentation.

The DDR repository is included in the R3 registry [14].

[1] Data Models and Metadata Policies
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/policies/curation/
[2] Core Scientific Data Model http://icatproject-contrib.github.io/CSMD/
[3] Curation and Publication of Simulation Data in DesignSafe https://doi.org/10.3390/publications7030051
[4] Best Practices to Enhance the Quality, Discoverability and Re-use Potential for Post-Event Reconnaissance Data
https://www.youtube.com/watch?v=xUyFJwZmyqM
[5] Disaster Risk Management Through the DesignSafe Infrastructure https://doi.org/10.1007/s13753-020-00320-8
[6] Publishing with the hybrid simulation data model in DesignSafe https://www.youtube.com/watch?v=iYzvYi-SY8Q
[7] Metadata Requirements Best Practices
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/best-practices/data-curation/
[8] Data Citation Policy https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/policies/publication/
[9] FAQs Publishing https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/faq/
[10] Data Usage Agreement
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/policies/publication/
[11] The Impact of Data Reuse https://www.designsafe-ci.org/rw/impact-of-data-reuse/
[12]Data Impact Policy https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/policies/publication/
[13]Data Marketing Best Practices
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/best-practices/data-publication/
[14] R3 Data Registry of Research Data Repositories https://www.re3data.org/repository/r3d100012308

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

14. Data reuse

R14. The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept.

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept 4.

Response:

Many data publications in the DDR are large and complex, and many span various years during which data producers release different components (e.g., experiments, components of longitudinal studies, versions). Example of such datasets can be found within the winners of the 2021 DesignSafe Dataset Award [1]. Their understandability is a concern that we are addressing through graphical interfaces, domain specific metadata, and documentation requirements.

The data models and schemas we use for metadata are designed and updated in concert with our core community, to make sure the terms used are relevant and current to the state of the practice in natural hazards research [2]. In turn, they are mapped to metadata standards for purposes of standardization and exchange [3]. The Fedora implementation also allows us to export data/metadata packages in a standard format [4].

When data are accessed, users can visualize all the metadata in the project landing pages in relation to the structure of the research project types. The interface design enables users to see the different data components of the project and their relations to the processes from which the data originated. A data graph – specifically useful for understanding big datasets encompassing multiple simulation runs, field missions or experimental tests - is provided to facilitate navigating the projects. To indicate the role of files in the midst of large numbers, we encourage users to use folder and file tags with terms specifically contributed by the community. Our intention is to facilitate tracing of the relationships between datasets by highlighting provenance. Our data models and subsequent publication policy [5] facilitate publishing datasets derived from analyses performed on original data, and to reference the subsets of data used in those analysis in the Referenced Data field. The interfaces have been evaluated by independent usability consultants both prior to and after they were released.

We provide DOIs to all published data to make these objects easier to find and reference in the future.
We require that users publish a data report along with their datasets. Those can have the form of readme files, data dictionaries, or white papers. We also ask users to add publications and other contextual information as Related Work. A metadata file is included for download with each published project. The metadata includes the description of the project as well as of each of the data components included in the package.

As described in R8, our Data Formats Policy [6] states that the DDR accepts all data formats users upload and publish with, and we deposit their datasets as is, although we do recommend that users deposit their data in open formats whenever possible. Our Accepted and Recommended File Formats Best Practices [7] also list the formats we

recommend, as well as curation primers. It is accepted (and for adoption, likely required) practice within the community to share data in some proprietary formats, primarily Matlab, Excel, and Labview. While proprietary, these are ubiquitous and the best formats for promoting reuse – and although they evolve, all are likely to have translators for the foreseeable future, as each has a dominant position in its respective market (we assume Matlab and Excel are well-known – Labview from National Instruments has dominant market share in segments of the computer-controlled instrumentation market). In addition, we allow users to publish when feasible both in proprietary and open formats. Within the broader DesignSafe workspace, we have available converters users can employ to move from older published formats to newer ones, and this set of tools will evolve with the formats.

Enabling data reuse is one of our main goals as a repository. We measure access and reuse to digital objects as much as practically possible and publish the results on the projects landing pages [8]. We sponsor community prizes for outstanding datasets in which both scientific merit and curation are evaluated by community experts [1], and highlight instances of reuse [9].

[1] DesignSafe Dataset Awards 2021 https://www.designsafe-ci.org/community/news/2021/march/dataset-awards/
[2] Data Models and Metadata Policies
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/policies/curation/
[3] Metadata Requirements Best Practices Metadata Dictionary
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/best-practices/data-curation/
[4] Fedora Import and Export Tools https://wiki.lyrasis.org/display/FEDORA5x/Import+and+Export+Tools
[5] Subsequent Publishing Policy
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/policies/publication/
[6] Data Formats Policy https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/policies/collections/
[7] Accepted and Recommended Data Formats Best Practices
https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/best-practices/
[8] Data Impact Policy https://www.designsafe-ci.org/rw/user-guides/curating-publishing-projects/policies/publication/
[9] The Impact of Data Reuse https://www.designsafe-ci.org/rw/impact-of-data-reuse/

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

TECHNOLOGY

15. Technical infrastructure

R15. The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept.

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept 4.

Response:

For the preservation environment itself, we use the Fedora Repository system, which is standards compliant and follows an upgrade path to maintain the standard [1]. Below the level of Fedora, we use the IBM Spectrum Scale Filesystem, implemented in a standard Linux distribution. The offline archive uses Quantum's StorNext software as a hierarchical storage manager. The front-end web interface to the repository is implemented via Python, Javascript, and React.js; versions of these tools are updated regularly to stay within fully supported versions.

In relation to planned infrastructure development DDR relies primarily on three TACC systems: the virtual machine cluster Roundup, the data collections system Corral [2], and the archival system Ranch [3]. All of these are shared-tenancy systems; the DDR is one of many tenants that make use of each of these pieces of infrastructure. Roundup uses VMWare and OpenStack on Dell servers to implement virtual machines, and DesignSafe and the DDR run in Docker containers hosted on these VMs. Licenses are renewed regularly, and the Roundup hardware is regularly augmented with purpose-bought servers as well as re-purposed servers from TACC's supercomputing resources. Corral is a large-scale data storage system, replaced every 5-6 years, providing a variety of file, object, and relational data services. The DDR makes use of the Spectrum Scale Filesystem, which is an asynchronously replicated instance between two sets of hardware. The current Corral is the fourth incarnation of the system. The existing hardware is from DataDirect Networks (DDN), with a planned replacement using IBM storage. The final system is the tape archive Ranch, used for offline backups of the primary online copies. This system is also replaced periodically and has been in continuous operation for 35 years. The current instance uses DDN storage for the first tier, and both Quantum and IBM tape libraries for the second tier. Both Ranch and Corral are periodically refreshed from a combination of cost recovery funds from customers and

University of Texas funding sources.

For the underlying systems upon which the repository operates, the public user guides document available software versions and capabilities. Internal, staff-only systems (using Redmine) track configuration management on all platforms.

The full repository has an extensive software stack, much of it is open source, such as Fedora, Linux, etc. Most of these are either well-supported externally (Linux), supported in-house (our portal APIs), or could be replaced by alternatives.

TACC maintains multiple, redundant 100Gbps network paths to Internet2 and 10Gbps paths to commodity internet network providers. The repository has never used more than a tiny fraction of available bandwidth.

As for our disaster and business continuity plan, the primary online storage is replicated between data centers, and a third offline copy is in a third data center. The second copy can be made the primary target if necessary. The front end is a set of containers with backed up images, allowing restoration on new hardware. These measures are only used in the event of corruption or disaster – the nature of the repository as a research resource allows us to tolerate small amounts of downtime for system maintenance, cut fibers, etc.; but in the event of a critical event the full repository could be reconstructed from offsite backups.

[1] Fedora Repository https://duraspace.org/fedora/
[2] Corral https://www.tacc.utexas.edu/systems/corral
[3] Ranch https://www.tacc.utexas.edu/systems/ranch

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

16. Security

R16. The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept.

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept 4.

Response:

The DDR exists within the larger TACC security context. Security is managed by TACC's Information Security Officer and his team with more than twenty full time systems administration staff. The complete description of TACC's security policies is available in our "System Security Plan." Since we do not make our security plan public, we have appended it in this application. Our environment is audited bi-annually by an outside auditor (most recently Crowe-Horvath) for compliance with NIST, FISMA, and HIPAA standards. The most recent audit (2020) found no outstanding issues.

Among the many controls – all direct system access has multi-factor authentication. Elevation of privilege requires an additional multi-factor authentication via RSA Token. All systems are patched on a regular basis, and relevant security lists are monitored daily for zero-day exploits. A firewall is in place to monitor every packet at the TACC network border, as well as an intrusion detection system running at line rate looking for patterns in network flows. All security-related information is logged via Splunk. Automated notifications reach operators who are on duty 24x7.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

APPLICANT FEEDBACK

Comments/feedback

These Requirements are not seen as final, and we value your input to improve the CoreTrustSeal certification procedure. Any comments on the quality of the Requirements, their relevance to your organization, or any other contribution, will be considered as part of future iterations.

Response:

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:
Good Application.
Documented web pages and additional pieces of information were given.

The applicant has made the amendments requested particularly on R3.