



# Fostering Geothermal Machine Learning Success: Elevating Big Data Accessibility and Automated Data Standardization in the Geothermal Data Repository

## Preprint

Nicole Taverna,<sup>1</sup> Jon Weers,<sup>1</sup> Scott Mello,<sup>1</sup> Adrienne Lowney,<sup>1</sup> Amber Mohammad,<sup>1</sup> and Sean Porse<sup>2</sup>

*1 National Renewable Energy Laboratory*

*2 U.S. Department of Energy, Geothermal Technologies Office*

*Presented at the 2024 Geothermal Rising Conference*

*Waikoloa, Hawaii*

*October 27-30, 2024*

**NREL is a national laboratory of the U.S. Department of Energy  
Office of Energy Efficiency & Renewable Energy  
Operated by the Alliance for Sustainable Energy, LLC**

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at [www.nrel.gov/publications](http://www.nrel.gov/publications).

Contract No. DE-AC36-08GO28308

**Conference Paper**  
NREL/CP-6A20-90400  
November 2024



# Fostering Geothermal Machine Learning Success: Elevating Big Data Accessibility and Automated Data Standardization in the Geothermal Data Repository

## Preprint

Nicole Taverna,<sup>1</sup> Jon Weers,<sup>1</sup> Scott Mello,<sup>1</sup> Adrienne Lowney,<sup>1</sup> Amber Mohammad,<sup>1</sup> and Sean Porse<sup>2</sup>

*1 National Renewable Energy Laboratory*

*2 U.S. Department of Energy, Geothermal Technologies Office*

## Suggested Citation

Taverna, Nicole, Jon Weers, Scott Mello, Adrienne Lowney, Amber Mohammad, and Sean Porse. 2024. *Fostering Geothermal Machine Learning Success: Elevating Big Data Accessibility and Automated Data Standardization in the Geothermal Data Repository: Preprint*. Golden, CO: National Renewable Energy Laboratory. NREL/CP-6A20-90400. <https://www.nrel.gov/docs/fy25osti/90400.pdf>.

**NREL is a national laboratory of the U.S. Department of Energy  
Office of Energy Efficiency & Renewable Energy  
Operated by the Alliance for Sustainable Energy, LLC**

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at [www.nrel.gov/publications](http://www.nrel.gov/publications).

Contract No. DE-AC36-08GO28308

**Conference Paper**  
NREL/CP-6A20-90400  
November 2024

National Renewable Energy Laboratory  
15013 Denver West Parkway  
Golden, CO 80401  
303-275-3000 • [www.nrel.gov](http://www.nrel.gov)

## NOTICE

This work was authored in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Geothermal Technologies Office. The views expressed herein do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at [www.nrel.gov/publications](http://www.nrel.gov/publications).

U.S. Department of Energy (DOE) reports produced after 1991 and a growing number of pre-1991 documents are available free via [www.OSTI.gov](http://www.OSTI.gov).

*Cover Photos by Dennis Schroeder: (clockwise, left to right) NREL 51934, NREL 45897, NREL 42160, NREL 45891, NREL 48097, NREL 46526.*

NREL prints on paper that contains recycled content.

# Fostering Geothermal Machine Learning Success: Elevating Big Data Accessibility and Automated Data Standardization in the Geothermal Data Repository

Nicole Taverna<sup>1</sup>, Jon Weers<sup>1</sup>, Scott Mello<sup>1</sup>, Adrienne Lowney<sup>1</sup>, Amber Mohammad<sup>1</sup>, and Sean Porse<sup>2</sup>

<sup>1</sup>National Renewable Energy Laboratory

<sup>2</sup>U.S. Department of Energy, Geothermal Technologies Office

## Keywords

*data, data standard, data pipeline, data lake, user experience, accessibility, gdr, data science, gis, geospatial, das, distributed acoustic sensing*

## ABSTRACT

The Department of Energy's (DOE's) Geothermal Data Repository (GDR) has implemented improvements to both its data lakes and its data standards and automated data pipelines. The GDR data lakes have reduced storage and compute-related barriers to using large geothermal datasets, enabling these large datasets to be accessed by anyone with a modern computer and internet access. More recently, the GDR has been working to further reduce barriers through streamlining the data intake process, educating users on the process and requirements, and helping users access data from the data lakes. These improvements have augmented the quantity of datasets the GDR is able to accept into its data lakes and have enabled users who are new to cloud tools to access these datasets more easily, overall increasing the accessibility of big geothermal data for use in machine learning and other projects. In addition, the GDR now has built-in data standards and pipelines for drilling data, geospatial data, and distributed acoustic sensing (DAS) data. These standardization efforts aim to enhance the real-world applicability of geothermal machine learning outcomes by improving the quality of training data. Specifically, through standardizing high-value datasets, the GDR is reducing project-specific data curation requirements, thus allowing more time for actual research. By automating this process, the burden of standardization is lifted from the user, ultimately increasing the availability of standardized data.

This paper provides an update on recent improvements made to the GDR's data lakes and automated data pipelines, including: (1) streamlining the data lake intake process, (2) better educating users on the process and requirements through a new data lakes page, (3) adding data lake direct access links to GDR data lake submission pages, (4) implementing a DAS data pipeline to convert DAS data uploaded in SEG-Y format to a standardized hierarchical data format v5 (HDF5), (5) extending this pipeline to encompass data in the GDR data lake, (6) adding metadata requirements for geospatial data, (7) making user interface/user experience (UX) enhancements to the data pipelines' documentation pages, and (8) improving the GDR's data standards and pipelines pages to better guide users in ensuring that their data is standardized by the GDR's automated data pipelines.

## **1. Introduction**

The U.S. Department of Energy’s (DOE’s) Geothermal Data Repository (GDR) serves as the repository and catalog for data generated by projects funded by the DOE Geothermal Technologies Office (GTO) (Weers et al. 2022). The GDR provides public access to geothermal datasets, which are consistently increasing in variety, size, and complexity. These datasets are becoming increasingly valuable for geothermal machine learning projects. To accommodate this, the GDR is improving its data lakes, implementing new robust data standards, and refining its data pipelines (Taverna et al. 2023(a,b), Weers et al. 2021). Additionally, improvements in user experience (UX) aim to make submitting, accessing, and utilizing datasets more convenient and efficient for researchers. These efforts collectively ensure that the GDR remains a crucial resource for advancing geothermal machine learning and data science projects.

### ***1.1 GDR Data Lakes***

As research datasets increase in size, the classic access model of downloading a dataset to a local compute resource becomes less feasible and can result in barriers to access and inefficiencies. Data lakes are a modern approach to data storage and access that provide a solution to these challenges. In the data lake model, especially large or complex (i.e., hierarchical) datasets are housed in a central, public-facing, cloud-based data store (Weers et al. 2021). The GDR data lakes not only enable users to parse or access data without needing to download the full dataset, but they also allow submitters to upload data without needing to rely on the traditional method of uploading data to the GDR submission form. Cloud-based data lakes have made big GDR data accessible to anyone with cloud access, eliminating the need for collaborators to have their own high-performance computing and big data storage solutions (Weers et al. 2021).

The GDR data lakes are currently home to more than 700 TB of public data, including distributed acoustic sensing (DAS), big geospatial data, raw magnetotellurics binary files, and more. Hosting the GDR’s big data in the GDR’s data lakes has greatly improved accessibility of these datasets and therefore the potential to use them in machine learning and data science workflows. Despite this, there have been challenges surrounding user understanding of data lake access, the data lake intake process, and data lake submission requirements. We have attempted to mitigate some of these challenges through streamlining the dataset intake process, better educating users on the process and requirements, and providing tools on the GDR submission page to ease access of GDR data lake data.

### ***1.2 Data Standardization and Automated Data Pipelines***

High-quality data is another key component of machine learning applications. High-quality geothermal datasets are characterized by reliable sensors or devices, frequent measurements, sufficient data points, comprehensive metadata, secure data storage, and effective data curation. Another aspect contributing to data quality is reusability, which can be improved through standardization. Standardizing data ensures consistency in formatting and content across similar datasets, reducing preprocessing requirements and ensuring that the dataset provides adequate information (Taverna et al. 2023(b)). Currently, the GDR has developed data standards for drilling data in its most common formats (Taverna et al. 2023(a)), DAS data in SEG-Y format (Taverna et al. 2023(b)) using PRODML (PRODML Work Group 2022) and the DAS RCN’s metadata standard (IRIS DAS RCN Metadata Working Group 2022), and geospatial metadata. Each

standard has an associated page on the GDR to provide additional information on how the standards work. These pages have recently undergone a UX review to improve their usability and accessibility.

The GDR has begun implementing automated data pipelines for certain high-value data types in order to take the burden of data standardization off of the user and ensure that the benefits of data standardization are felt more widely across the GDR (Taverna et al. 2023(b)). These data pipelines align with the data standards. The GDR previously only had one automated data pipeline for drilling data (Taverna et al. 2023(a)) but has since expanded this to include DAS and geospatial data.

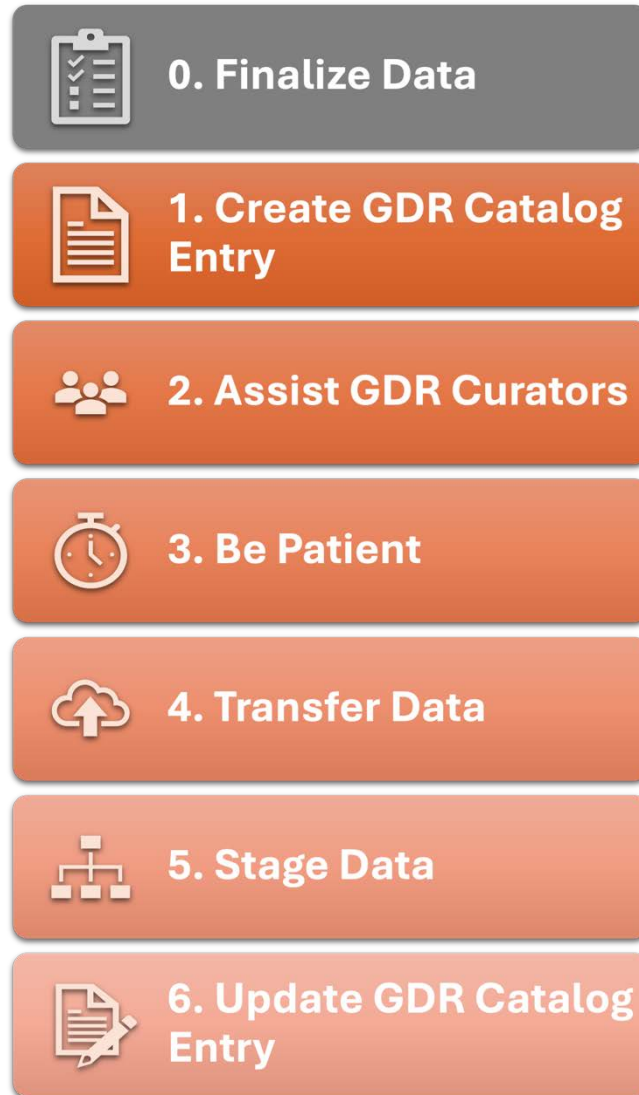
## 2. Streamlining the Data Lake Intake Process

Previously, the GDR data lake submission process was characterized by a lack of consistency, with only a few individuals possessing partial knowledge of the process, confusing conditional and inconsistent requirements, and a heavy reliance on ad hoc communication from curators. This resulted in a lack of standardization in user experience and requirements, as well as an outward lack of clarity on how long the process takes and what it entails. The GDR has worked to remedy this problem through streamlining the data lake submission process.

As a reminder, regular datasets should be submitted as normal via the GDR submission form. However, especially large or complex datasets should follow these steps, which are also condensed into the graphic in Figure 1:

1. Create a GDR catalog entry with associated metadata and supporting files. Supporting files may include related publications, readme files, GitHub repositories, a few sample files, or any other information that may help to give context to our curators. Do not try to upload the entire dataset to the submission form. Instead, contact our curation team at [GDRHelp@ee.doe.gov](mailto:GDRHelp@ee.doe.gov) and request a data lake resource.
2. Assist GDR curators with public datasets program application requirements. Some data lake datasets may be eligible for free hosting and/or registration in the data catalogs of our cloud partners, such as the [Amazon Sustainability Data Initiative](#). This may include filling out an additional questionnaire to justify the value of your dataset, helping to create a tutorial Jupyter notebook with examples for accessing and using the data, and providing any additional requested resource metadata.
3. Wait to hear back about approval from third-party public datasets programs. The review process is currently quarterly, so please be patient and plan on a few months for this step.
4. Assist data curators with the data transfer process either by making the data available via Globus, FTP, Google drive, or another web-based file-sharing service or by moving the data directly into our staging bucket.
5. When the data have been properly staged and organized, the GDR curation team will ask you to review the data one more time for organization, naming, and completeness. Data will be considered permanent and final after this step unless stated otherwise.
6. Assist curators with any final metadata tweaks. If you are publishing a journal article related to the data, share a preprint with the GDR Team and follow up when the article is published.

Please note that the GDR does not validate or peer review datasets. **All data should be finalized, validated, and peer reviewed by the project team prior to submission to the GDR.** Datasets published to the GDR are considered scientific records, and their data resources are expected to remain unchanged in the same location permanently. Incremental additions can be easily added to existing datasets, including the addition of new versions of previously published data. However, publishing changes to existing data or other major updates to datasets—including data in data lakes—often requires creating an entirely new version of the dataset.



**Figure 1: Graphical representation of the streamlined data lake intake process.**

This new streamlined process standardizes the user experience and makes the intake process more efficient. This has been evident in our ability to scale up data lake data intake to meet the growing demands of big geothermal data storage. The 475 TB of DAS data currently available through the GDR is a great example of this achievement.



### 3. User Education and Support

In addition to lacking consistency, the data lake intake process has also lacked transparency. User feedback has relayed that there is a lack of understanding surrounding the submission requirements, the submission timeline, and the GDR’s data posterity requirements. To mitigate this issue, the GDR has decided to focus on more proactive education and support. This has included creating a [new page on the GDR](#) to describe the data lake intake process in detail, including the graphic in Figure 1, along with more in-depth information about what to expect during the data lake intake process. This has also included work to publish and present on these requirements—including this paper. These steps allow the GDR to clearly communicate the requirements and expectations to potential submitters prior to initiating the submission process.

The impact of this proactive education approach has been observed through an overall decrease in requests that conflict with our policies, an improvement in organization of new submissions, and a reduction in proliferation of largely duplicative versions of data in our data lakes. These factors combined result in higher quality data lake submissions overall.

### 4. Direct Access Links for Data Lake Data

Users have frequently asked questions surrounding how to efficiently access data from GDR data lakes. In order to make this information more readily discoverable, the GDR has added direct access links to the submission form for data lake resources. These direct access links allow users to easily copy and paste access commands for accessing the data using the Amazon Web Service Command Line Interface (AWS CLI). An example of a direct access link is shown in Figure 2.



**Figure 2: Example of a direct access link associated with a GDR data lake resource. Users can use the icon to the right to quickly copy the command so that they may paste it into their terminal for AWS CLI access to the data.**

Additionally, the GDR has listed multiple ways to interact with data in data lakes on the [Data Lakes page](#), accessible from the Data drop-down at the top of every page. This list, complete with examples, illustrates how to access GDR data through:

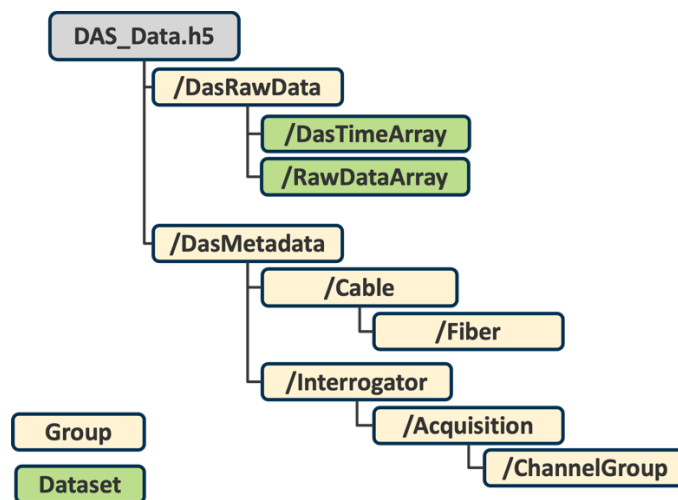
- Jupyter notebooks
- Native cloud query engines like Athena (AWS) and Google Earth Engine
- Native cloud access direct to the public data lakes
- The GDR Data Lake Viewer, a simple user interface that allows users to explore data lakes through their web browser
- Native cloud command line tools, such as the AWS CLI example above
- Mounting the data as a local read-only drive in a cloud-built computer cluster (requires a cloud account in the same availability zone as the data lake.)



## 5. Implementation of DAS Data Pipeline

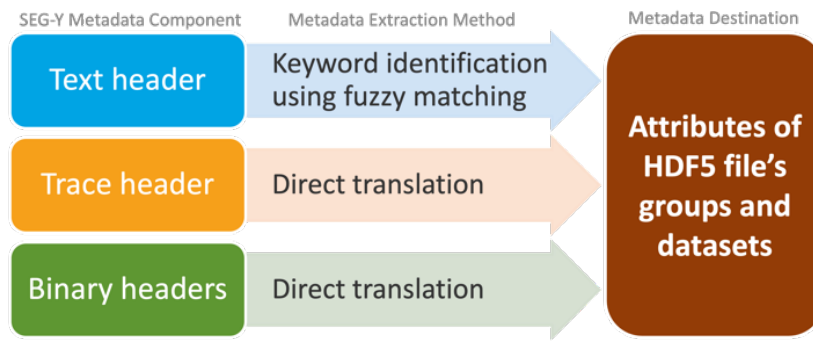
To maximize the utility of the large quantity of DAS data stored in the GDR, a data pipeline has been implemented to convert DAS data in nonstandard SEG-Y format to a standardized hierarchical data format (HDF5). The data standard combines aspects of the PRODML raw DAS data format (PRODML Work Group 2022) with the DAS RCN Metadata Working Group’s metadata standard (IRIS DAS RCN Metadata Working Group 2022).

This hybridized approach is described in Figure 3 and is described in more detail in Taverna et al. 2023(b). The file structure is hierarchical, broken into two main groups: one for raw DAS data, and one for DAS metadata. The DAS raw data group includes two datasets—one containing a time array, and another containing a raw data array. The metadata group includes two types of subgroups: one for each cable’s metadata that contains subgroups for each fiber’s metadata, and one for each interrogator’s metadata, which is further broken down into acquisition and channel-specific metadata. While the DAS RCN recommends that metadata be stored in a separate JSON file, the GDR standard attempts to store as much of it as possible within the HDF5 file to keep the metadata packaged with the data. Channel location information, however, is stored in a separate channel map file to allow for simplified updates.



**Figure 3: GDR DAS data standard file structure. The GDR standard is HDF5-based, combining aspects of the PRODML format with the DAS RCN Metadata Working Group’s metadata standard.**

The DAS data pipeline generally operates as described in Figure 4. Metadata are stored in three main locations within a SEG-Y file: The text header, trace header, and binary headers. Because metadata in the text header are not standardized by nature, fuzzy matching is used to identify keywords in the text header, and these keywords are used to extract desired information. Metadata are standardized by nature in the trace and binary headers, so direct translation may be used to extract this information. Extracted metadata are then stored in the associated attributes of an HDF5 file’s groups and datasets. This pipeline is not fully automated yet because of the massiveness of DAS datasets, the diverse ways that GDR DAS datasets are organized, and the frequently incomplete DAS metadata packages.



**Figure 4: Graphic depicting the general process followed by the DAS data pipeline. Metadata are extracted using a combination of direct translation and fuzzy matching and stored in the associated attributes of the standardized HDF5 file’s groups and datasets.**

While we are retroactively applying this data pipeline to existing DAS datasets in the GDR, many of these existing datasets have incomplete metadata packages. This is because the most common GDR DAS data format, SEG-Y, is a file format intended for conventional seismic data, not DAS data, meaning that it is not compatible with all of the metadata required for robust DAS data analysis. On top of that, there was no universally agreed on DAS metadata standard until the recent work of the DAS RCN Metadata Working Group (2022). In an attempt to maximize the amount of metadata added to the standardized DAS data files, a metadata template was created to allow data owners to provide additional metadata. GDR curators are actively reaching out to legacy DAS data owners requesting additional metadata via this form.

At the time of this publication, 16.5% of the approximately 287 TB of DAS data in the GDR has been standardized using this method. The remaining data are in formats (non-standardized HDF5, TDMS, and zipped directories) that are not yet compatible with our data pipeline. Future work will investigate the merit of adapting our pipeline to additionally standardize one or more of these formats.

## 6. Extension of Data Pipelines for GDR Data Lake

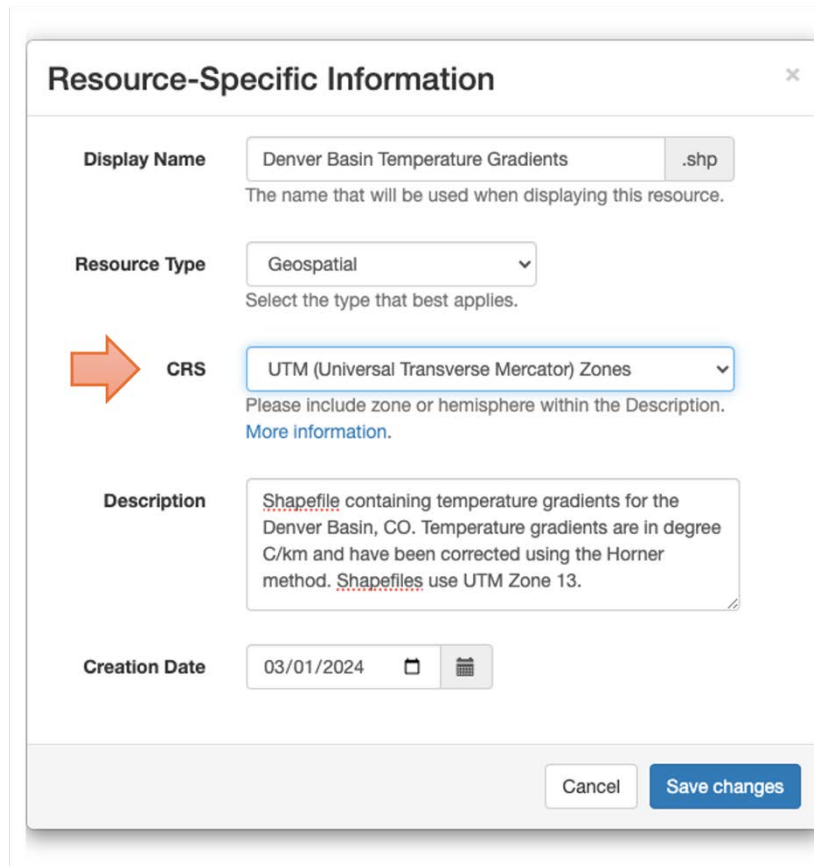
The DAS data pipeline is the first to be adapted to work for data in the GDR data lakes. Because DAS data are often on the order of tens of terabytes in size (Trainor-Guitton et al. 2022), they are almost exclusively stored in the GDR data lakes. This means that, in order to be of maximum utility, the data pipeline needed to be adapted to work on data in the data lakes. This necessitated adjustment of the data detector script to detect DAS data files in specific directories in the cloud as opposed to on the GDR servers.

This adaptation will streamline the development of future data pipelines that are compatible with big data in the data lake, which is arguably where data standardization provides the most value. Big datasets require more time-consuming data curation; therefore, standardization of big data results in more significant time savings through reducing the curation requirements (Taverna et al. 2023(a)).

## 7. Added Metadata Requirements for Geospatial Data

Current storage models often lack the necessary metadata for enabling thorough analyses and reproducibility. Consequently, GDR submissions of geospatial data from GIS nonexperts have frequently lacked complete geospatial metadata (Taverna et al. 2023(b)). To address this problem, a comprehensive review was conducted of proper metadata to be archived alongside geospatial datasets (Taverna et al. 2023(b)). It was decided that the most crucial piece of geospatial metadata is the coordinate reference system (CRS), and hence a required input for CRS has been added to the resource-specific metadata for files with geospatial file extensions (Figure 5). This required input ensures that the CRS is added for every geospatial data file added to the GDR and recommends additional CRS-specific metadata to add to the resource description.

Beyond the CRS, the GDR geospatial metadata standard provides suggestions of which other geospatial metadata is useful to include in your GDR submission, and where to include it in the GDR submission form. These suggestions may be found in Taverna et al. 2023(b) or on the GDR's new geospatial data standard page: [https://gdr.openei.org/geospatial\\_data\\_standard](https://gdr.openei.org/geospatial_data_standard). The GDR has already seen an improvement in the quality of its geospatial datasets in the short time since this standard and the additional required input have been implemented.



The image shows a web form titled "Resource-Specific Information" with a close button (X) in the top right corner. The form contains several fields:

- Display Name:** A text input field containing "Denver Basin Temperature Gradients" and a file extension ".shp" in a separate box. Below it is the text: "The name that will be used when displaying this resource."
- Resource Type:** A dropdown menu with "Geospatial" selected. Below it is the text: "Select the type that best applies."
- CRS:** A dropdown menu with "UTM (Universal Transverse Mercator) Zones" selected. An orange arrow points to this field. Below it is the text: "Please include zone or hemisphere within the Description. [More information.](#)"
- Description:** A text area containing the text: "Shapefile containing temperature gradients for the Denver Basin, CO. Temperature gradients are in degree C/km and have been corrected using the Horner method. Shapefiles use UTM Zone 13."
- Creation Date:** A date input field showing "03/01/2024" with calendar icons.

At the bottom right of the form are two buttons: "Cancel" and "Save changes".

**Figure 5: Image of new CRS requirement for geospatial data files uploaded to the GDR. Note the additional CRS-specific metadata suggestion that appears below the CRS field.**

## 8. UX Enhancements

In order to improve the UX associated with the GDR data standards and pipelines, a UX review was conducted on the existing GDR data standard and pipeline pages. This heuristic review primarily focused on the following user-centric design principles: user needs and goals, accessibility, simplicity and clarity, responsive design, feedback and performance, and trust and credibility. This was achieved through revisiting the GDR's data standard and pipeline pages, identifying bugs, confusing wording, and non-intuitive designs; and providing suggestions on how to improve functionality, clarity of information, intuitiveness of navigation, and more.

In this heuristic review, several design principles are recommended to enhance the user interface and functionality of a data table. Key principles include improving navigation by making table headers sticky to reduce scrolling lag and adding a keyword search feature for efficient data filtering. Clear organization is suggested by adding distinct headers for different data sections and creating separate columns for specific data categories. To improve readability, long descriptions should be truncated with an option to expand for more details, and sorting capabilities should be added for key columns. User interaction can be enhanced with icons indicating required fields and maintaining row highlights on hover. Accessibility improvements include ensuring color contrast.

## 9. Improvements in Data Standards and Pipelines Pages

The input gathered from the UX review was then combined into new and improved data standard and pipeline pages. The key takeaways were then applied to the new data standard and pipeline pages to ensure that they adhered to the same design principles. This included enhancements to design and layout, in addition to pipeline guides and user instructions (i.e., tips to make sure your data are standardized, compatible formats, helpful resources, and how it works). These updates may be found on the data standards and pipeline pages here: <https://gdr.openei.org/standards>. While it is too soon to tell, the GDR team anticipates that these new pages will have positive impacts on the amount of data that can be automatically standardized, the use of standardized data, and the overall support for the GDR data standards and pipelines.

## 10. Discussion: Real-World Applicability and Impact on Research

The Geothermal Operational Optimization with Machine Learning project exemplifies how emphasizing data quality through best practices in data curation can significantly enhance the success of a machine learning project (Taverna et al. 2022). Having access to structured and standardized data streamlines data handling by reducing time spent on dataset preparation, enabling efficient assembly of multiple datasets. Having access to especially large or complex datasets through data lakes eases data access, reducing barriers surrounding downloading, uploading, and working with the data. Combined, these features lessen the data curation requirements associated with a project, preserving more time for machine learning and data science experiments.

While it is too soon to see the scale of the impact of the GDR's data standardization efforts, we hope to see growth in the number of geothermal projects using standardized GDR data in their machine learning and data science workflows. We have observed the impact of the GDR data lakes through the sheer number of downloads of our data lake resources.

Table 1 lists the top ten most downloaded datasets of all time, with six of them being data lake resources. This trend is partly due to the nature of data lakes, which often contain especially large and complex datasets comprising numerous files accessed directly from the data lake (i.e., using files in situ instead of a one-time download). Additionally, this highlights the popularity of GDR data lake resources, which significantly enhance the accessibility of extremely large datasets. Notably, the first data lake resource was introduced nine years after the first conventional data resource, which was submitted in 2012.

**Table 1: Top 10 most downloaded data submissions of all time as of August, 2024, including information on number of downloads, year published, and dataset storage type.**

Rank	Downloads	Year Published on GDR	Title	Dataset storage type
1	5,845,469	2021	PoroTomo Natural Laboratory Horizontal and Vertical Distributed Acoustic Sensing Data	Data lake
2	1,561,471	2021	Brady’s Geothermal Field - March 2016 Vibroseis SEG-Y Files and UTM Locations	Data lake
3	1,089,412	2021	Brady’s Geothermal Field Nodal Seismometer Data	Data lake
4	728,186	2014	Evaluation Data of a High Temperature COTS (Commercial Off-the-Shelf) Flash Memory Module (TI SM28VLT32) for Use in Geothermal Electronics Packages	Conventional
5	434,950	2021	Brady’s Geothermal Field DASH Resampled in Time	Data lake
6	309,362	2023	EGS Collab Experiment 1: Continuous Active-Source Seismic Monitoring (CASSM) Data	Data lake
7	28,441	2014	Project HOTSPOT: Kimama Well Borehole Geophysics Database	Conventional
8	22,472	2023	Imperial Valley Dark Fiber Project Continuous DAS Data	Data lake
9	18,126	2014	Project HOTSPOT: Kimberly Well Borehole Geophysics Database	Conventional
10	12,807	2013	Newberry EGS Demonstration Well 55-29 Stimulation Data	Conventional

This statistic also gives context to the plot in Figure 6, which shows cumulative GDR downloads per fiscal year since the GDR was rolled out more than 10 years ago. Because data lake downloads make up most of the FY24 downloads, it is apparent that the exponential growth in downloads over time would not be possible without the implementation of GDR data lakes.



**Figure 6: Cumulative GDR downloads over time, by fiscal year, as of August 2024. This graph demonstrates exponential growth in the number of downloads per fiscal year, enabled by the GDR data lakes (Graphic by Joelynn Schroeder and Dominique Barnes, NREL, 2024).**

## 11. Conclusion

In conclusion, the GDR has made significant advancements in its data lakes, data standards, automated data pipelines, and associated pages to educate users. These enhancements have reduced barriers related to storage, access, and computation, enabling broader access to large geothermal datasets. Recent efforts to streamline the data intake process, educate users, and improve data accessibility have increased the number of datasets the GDR can accept into its data lakes and made it easier for new users to access the data using cloud tools. Introducing standardized pipelines for drilling, geospatial, and DAS data has improved the quality of training data, thereby enhancing the real-world applicability of geothermal machine learning outcomes. Automating

these processes has increased the availability of standardized data, reducing the burden of project-specific data curation on users and allowing more time for research. Overall, these improvements have significantly increased the accessibility and usability of big geothermal data for various projects.

We encourage geothermal data users to leverage the GDR's enhanced data lakes, standardized datasets, and streamlined data pipelines; and to provide feedback on their experience, including both success stories and shortcomings. Your active participation and feedback are crucial for driving innovation and ensuring that future improvements to the GDR align with the needs and wants of the geothermal community and stakeholders. We would love for you to join us in this initiative to make geothermal data more accessible, efficient, and impactful for the entire community.

## **Acknowledgement**

This work was authored in part by the National Renewable Energy Laboratory (NREL), operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DEAC36-08GO28308 with funding provided by U.S. Department of Energy Office of Energy Efficiency and Renewable Energy (EERE) Geothermal Technologies Office (GTO). The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains, and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

We'd like to recognize the National Geothermal Data System content models as the catalyst for the work on data standardization. We'd also like to acknowledge Aleksei Titov, Kurt Feigl, Herb Wang, Kathleen Hodgkinson, Rob Porritt, and the rest of the DAS RCN Metadata working group for their input on and inspiration for our DAS data standard. In addition, we thank Meghan Mooney at NREL for her review of and suggestions around the geospatial metadata and data standards proposed in this paper. And lastly, we acknowledge OpenAI's ChatGPT 4 for its assistance in brainstorming and proofreading.

## **REFERENCES**

- The Geothermal Data Repository (GDR). <https://gdr.openei.org/home>.
- IRIS DAS Research Coordination Network (RCN) Metadata Working Group. "Distributed Acoustic Sensing (DAS) Metadata Model." Whitepaper (2022). [https://github.com/DAS-RCN/DAS\\_metadata](https://github.com/DAS-RCN/DAS_metadata).
- PRODML Work Group, "PRODML Technical Reference Guide." Energistics, v2.2 (2022). <https://www.energistics.org/prodml-data-standards>.
- Taverna, N., Weers, J., Porse, S., Anderson, A., Frone, F., Holdt, E. 2023. "An Update on the Geothermal Data Repository's Data Standards and Pipelines: Geospatial Data and Distributed Acoustic Sensing Data" *GRC Transactions, Vol. 47* (2023).



- Taverna, N., Weers, J., Huggins, J., Porse, S., Anderson, A., Frone, F., Scavo, R.J. 2023. “Improving the Quality of Geothermal Data Through Data Standards and Pipelines Within the Geothermal Data Repository.” *Proceedings of the 48th Workshop on Geothermal Reservoir Engineering*, Stanford University, Stanford, CA (2023).
- Taverna, N., Buster, G., Huggins, J., Rossol, M., Siratovich, P., Weers, J., Blair, A., Siega, C., Mannington, W., Urgel, A., Cen, J., Quinao, J., Watt, R., and Akerley, J. “Data Curation for Machine Learning Applied to Geothermal Power Plant Operational Data for GOOML: Geothermal Operational Optimization with Machine Learning.” *Proceedings of the 47th Workshop on Geothermal Reservoir Engineering*, Stanford University, Stanford, CA (2022).
- Trainor-Guitton, W., Martin, E. R., Rodríguez Tribaldos, V., Taverna, N., and Dumont, V. “Distributed sensing and machine learning hone seismic listening.” *Eos*, 103 (2022).
- Weers, J., Anderson, A., and Taverna, N. “The Geothermal Data Repository: Ten Years of Supporting the Geothermal Industry with Open Access to Geothermal Data.” *GRC Transactions*, Vol. 46 (2022).
- Weers, J., Porse, S., Huggins, J., Rossol, M., and Taverna, N. “Improving the Accessibility and Usability of Geothermal Information with Data Lakes and Data Pipelines on the Geothermal Data Repository.” *GRC Transactions*, Vol. 45 (2021).