An Update on the Geothermal Data Repository's Data Standards and Pipelines: Geospatial Data and Distributed Acoustic Sensing Data

Preprint

Nicole Taverna,[1] Jon Weers,[1] Sean Porse,[2]
Arlene Anderson,[2] Zachary Frone,[2] and Emily Holt[1]

1 National Renewable Energy Laboratory
2 U.S. Department of Energy

Presented at the Geothermal Rising Conference
Reno, Nevada
October 1-4, 2023

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.

Contract No. DE-AC36-08GO28308

Conference Paper
NREL/CP-6A20-86935
October 2023

An Update on the Geothermal Data Repository's Data Standards and Pipelines: Geospatial Data and Distributed Acoustic Sensing Data

Preprint

Nicole Taverna,[1] Jon Weers,[1] Sean Porse,[2]
Arlene Anderson,[2] Zachary Frone,[2] and Emily Holt[1]

1 National Renewable Energy Laboratory
2 U.S. Department of Energy

Suggested Citation
Taverna, Nicole, Jon Weers, Sean Porse, Arlene Anderson, Zachary Frone, and Emily Holt. 2023. *An Update on the Geothermal Data Repository's Data Standards and Pipelines: Geospatial Data and Distributed Acoustic Sensing Data: Preprint.* Golden, CO: National Renewable Energy Laboratory. NREL/CP-6A20-86935. https://www.nrel.gov/docs/fy24osti/86935.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.

Contract No. DE-AC36-08GO28308

Conference Paper
NREL/CP-6A20-86935
October 2023

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • 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.

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

An Update on the Geothermal Data Repository's Data Standards and Pipelines: Geospatial Data and Distributed Acoustic Sensing Data

Nicole Taverna[1], Jon Weers[1], Sean Porse[2], Arlene Anderson[2], Zachary Frone[2], and Emily Holt[1]

[1]National Renewable Energy Laboratory

[2]Geothermal Technologies Office, U.S. Department of Energy

Keywords

data, data standard, data pipeline, machine learning, data curation, gdr, data science, gis, geospatial data, das, distributed acoustic sensing data

ABSTRACT

The Department of Energy's (DOE) Geothermal Data Repository (GDR) team has implemented or is currently implementing data standards and automated data pipelines for the following geothermal data types: 1) drilling data, 2) geospatial datasets, and 3) Distributed Acoustic Sensing (DAS) data. These data standards and pipelines are intended to improve the real-world applicability of geothermal machine learning outputs through improving the quality of data. More specifically, through standardizing high-value datasets, the GDR is reducing project-specific data curation requirements, allowing more time to be spent on actual research. By automating this process, the burden of standardization is taken off of the user, overall increasing the availability of standardized data. This paper provides an update on the GDR's transition toward data standardization through automated data pipelines and calls for feedback from the community on how the GDR team can improve this process.

1. Introduction

The U.S. Department of Energy (DOE) Geothermal Data Repository (GDR) is 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. At the same time, these datasets are growing in value to geothermal machine learning projects. That considered, the GDR is constantly aiming to improve the convenience and efficiency of using its datasets in geothermal machine learning projects (Taverna et al., 2023).

High-quality data is essential for achieving high-quality results in machine learning applications. In the context of geothermal projects, the importance of data quality has been increasingly recognized. The Geothermal Operational Optimization Using Machine Learning (GOOML) project serves as an example of the positive outcomes that can be achieved with high-quality data.

The project utilized large amounts of geothermal power plant operational data to optimize power generation. Data curation played a crucial role in the success of the GOOML project, following a process that involved data acquisition, digestion, transformation, quality assurance, and utilization in machine learning algorithms. This iterative process focused on improving data quality rather than the more traditional approach of tuning model parameters, with the goal of enhancing the real-world applicability of geothermal machine learning projects. The GOOML data curation process emphasizes a data-centric approach, recognizing the critical role of high-quality data in project success (Taverna et al., 2022).

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. When submitting data to the GDR, preferred formats are those that support the highest reusability. While the GDR accepts a variety of file formats, it encourages the use of structured and standardized data whenever possible (Taverna et al., 2023). This tier of data has traditionally included standardized formats like Excel, CSV, XML, RDF, JSON, and others, promoting reusability and facilitating efficient data analysis (Tier 3 in Figure 1). Here we expand upon the existing data tiers to include a fourth, which is best for large or complex datasets (Tier 4 in Figure 1). This tier is not only structured and standardized, but is also cloud-optimized, offering advantages such as improved computational performance, storage cost efficiency, and scalability, especially when used in the cloud.

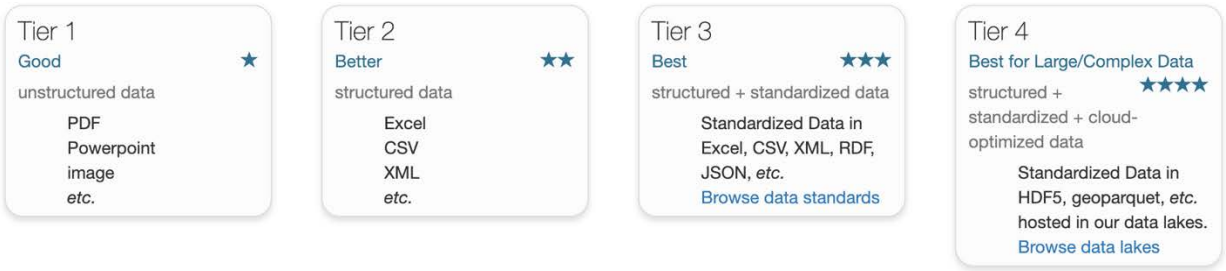

Figure 1: Graphic describing the GDR's guideline for preferred data formats. In this guideline, Tier 4: structured + standardized + cloud-optimized data is considered best because it maximizes reusability, while also enhancing cloud performance and efficiency.

1.1 Data Standards and Pipelines

Data standards provide data type-specific guidelines on contents, metadata, and format to help users submit data that falls under Tier 3 or Tier 4 in Figure 1. They can be used to advise data collection, or to reformat data after collection, with the goal of maximizing usability for future research. A small portion of the burden of meeting data standards falls on the submitter. For example, the submitter must structure the contents to some extent, provide complete metadata, and use a digitized, machine-readable format for their data. However, where automated data pipelines exist, submitters do not need to manually reformat or restructure their data to meet GDR's exact standards.

Automated data pipelines automatically recognize certain types of datasets, and then convert them into a standardized format while also preserving the original data file (Figure 2). This means that, if a data pipeline for that data type has been implemented in the GDR, researchers who produce the data can use whichever digitized and machine-readable data format they prefer internally, upload their data in that format, and have it be automatically standardized. This shift takes the majority of the burden of data standardization off the user and project teams, allowing more project resources to be used on research and development activities, and increasing the availability of standardized geothermal data available through the GDR.

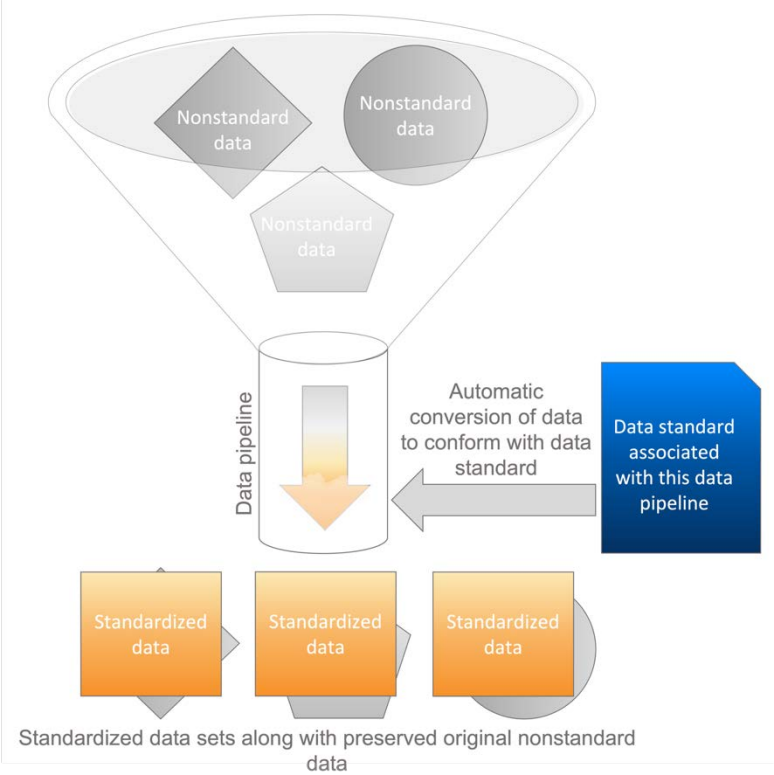

Figure 2: Graphic depicting how GDR data pipelines work. Nonstandard data is uploaded to the GDR and is funneled through a data pipeline that standardizes the data in accordance with the associated data standard, while preserving the original nonstandard data (Taverna et al. 2023).

1.2 Existing Data Standards and Pipelines in the GDR

In the past, the National Geothermal Data System (NGDS) content models (NGDS 2013) were used for data standardization. These models provided standardized templates in Excel and XML formats, but they placed the burden of standardization on the data submitter, which proved to be challenging and time-consuming, limiting adoption. Furthermore, the NGDS content models have limitations in capturing time-series data, big data, and non-tabular data, which are becoming more common in geothermal projects. As the datasets submitted to the GDR grow in variety, size, and complexity, a more robust approach to data standardization is needed to overcome these limitations and support a wider range of data types and formats. Therefore, the GDR is moving away from this model and towards one centered on automated data pipelines (Taverna et al., 2023).

The first automated data pipeline implemented within the GDR was for drilling data. This pipeline currently supports and is capable of processing data from Pason (Pason Systems Corp., 2023) and RigCloud (Nabors Industries Ltd., 2021) drilling data platforms in Excel and CSV formats and may, in the future, be amended to standardize drilling data from other sources and in other formats as well. The pipeline recognizes the platform-specific field names and units and converts them to the standard field names and units in CSV format. The standard additionally includes the RIMBase (Infostat, 2023) drilling data platform field names (Taverna et al., 2023).

When uploading a drilling dataset to the GDR, the submitter adds metadata, uploads a file containing drilling data in either Pason or RigCloud output formats, and saves or submits as normal. The data pipeline then automatically detects that the file contains drilling data, converts it into the data standard, and generates an additional file. Initially, the standardized file appears grayed out with a note indicating it is auto generated. After processing, (dependent on the file size, but usually after a few minutes), the standardized file becomes available for download alongside the original file, preserving both versions. This process is discussed in more detail and with figures in Taverna et al., 2023.

1.3 New Data Standards and Pipelines added to the GDR

The GDR team prioritized new data standards and pipelines based on anticipated awards, high-demand datasets, and feedback from the community. Internal brainstorming sessions were conducted, and an internal survey was administered to gather input on improvements desired for the GDR. Discussions were also held with DOE GTO to align efforts with upcoming projects. The results were presented at the Stanford Geothermal Workshop in February 2023, and another survey was administered to gather feedback from attendees. The final resulting priority new data pipelines and standards for 2023 included auto-detection and complete metadata requirements for submission of GIS data and a pipeline for converting non-standard Distributed Acoustic Sensing (DAS) data into a standard format.

1.3.1 Geospatial Data

Missing metadata impedes comprehensive analysis and reproducibility. High quality geospatial studies require complete metadata packages to enable precise mapping, explanation of potentially unexpected anomalies in the data, and uncertainty quantification. Current storage models often lack the necessary metadata for enabling thorough analyses and reproducibility. Consequently, data submissions from GIS non-experts frequently result in incomplete metadata.

Currently, the GDR encourages, but does not require any GIS-specific metadata. As a result, there are numerous GIS data files in the GDR with incomplete metadata. For example, many GDR submissions do not include a coordinate reference system (CRS). There have been efforts to try to identify the correct CRS for some of the high-value GIS datasets, but the process proved to be extremely time-consuming and introduced significant uncertainty. Based on this, it has been determined that the best solution for ensuring reproducibility and reusability for geospatial data is to require that crucial metadata are included upon upload of these files.

More recently, there has been a rise in the relevance of high-resolution geospatial data (i.e., GeoDAWN LiDAR data). These data are frequently on the order of terabytes in size, causing storage and access of these data using conventional approaches (i.e., storing in shapefiles and

opening in QGIS) to be rather inefficient. This lends to the need for additional data standards specific to big geospatial data, which are more focused on cloud optimization and ease of parsing using open-source tools such as python plotting libraries (e.g., matplotlib).

1.3.2 Distributed Acoustic Sensing Data

Distributed Acoustic Sensing (DAS) data has diverse applications, enabling the tracking of movement and activity in infrastructure and the discrimination of different sources of vibrations. DAS data is also employed in subsurface investigations, detecting earthquakes and characterizing subsurface structures. DAS, especially when combined with machine learning, offers valuable insights into subsurface phenomena and contributes to scientific and societal advancements (Trainor-Guitton et al., 2022), labeling it as a high-value data type within the GDR. Given the massiveness of DAS data, its potential for contributing to scientific discovery would be greatly augmented if efficient automated analysis (i.e., edge computing) of DAS data was made more convenient and scalable. Adopting a standardized format for DAS data helps to achieve this.

DAS uses an optical fiber to measure strain induced by elastic waves, providing continuous vibration sensing. DAS acts as an array of sensors along the fiber, measuring strain over discrete intervals. Unlike traditional seismic sensors, DAS captures directional strain components over a length, collects multiple channels simultaneously, has customizable acquisition parameters, and experiences variations in sensitivity due to installation environments. While DAS data is often used for seismic monitoring, standard seismic metadata (e.g., SEED) and file formats (e.g., SEGY) are ill-suited for DAS data due to their inability to accommodate acquisition parameters and handle large data volumes (IRIS DAS RCN Data Management Working Group, 2022).

Within DAS, a fiber optic cable is deployed in a well, like in Figure 3, or horizontally in a trench. The cable is composed of many fibers, which are themselves composed of many channels. The cable is hooked up to an interrogator unit which measures variations in back-scattered light caused by vibrations in the fiber and stores the accompanying raw data. This raw data is usually processed using data reduction, frequency-based filtering, or transformations. Applications can then derive information from either the raw data, the processed data, or both. These datasets are commonly on the order of terabytes in size, making traditional file storage and transfer challenging. In addition, the lack of standardized metadata and data formats mean that a significant amount of time is spent manually reformatting data and tracking down missing pieces of metadata, which can be costly and introduce errors (PRODML Work Group, 2022).

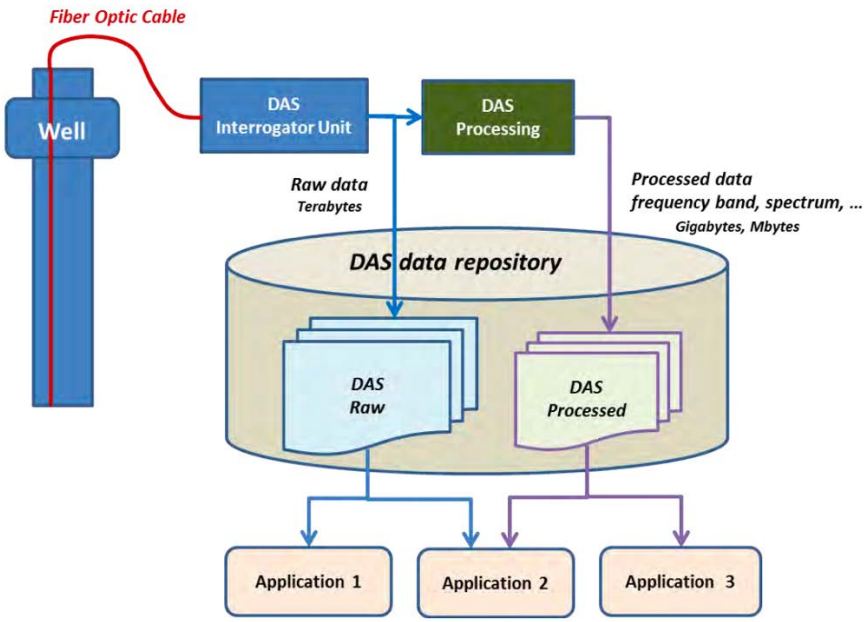

Figure 3: DAS data acquisition, processing, and storage overview (PRODML Work Group, 2022)

1.4 GDR Data Lakes

Very large datasets such as high-resolution geospatial data or DAS data are stored in the GDR data lakes rather than the traditional access model of downloading datasets for local use. Data lakes enable interaction with the data directly within the data lake. Researchers can send their research questions to the data lake and receive the answers without the need for data transfer. This can be achieved through encapsulating research questions in modular code or setting up a server or cluster of servers in a connected cloud environment. By eliminating the requirement for large data transfers, the data lake approach accelerates research timelines and reduces costs. Furthermore, the centralized nature of the data lake ensures consistent and equal access to the dataset for all collaborators, removing the need for data transfers between partners and reducing the risk of data corruption. Cloud-based data lakes are accessible to anyone with cloud access, eliminating the need for collaborators to have their own high-performance computing (HPC) and big data storage solutions (Weers et al., 2021).

2. Geospatial Data Standard

Since there is already an existing metadata standard for geospatial data (ISO 19115-1), the GDR's geospatial metadata standard is based on this. There is less consensus on preferred formats, so here we investigate some of the commonly used formats, suggest a format for very large geospatial data, and discuss updates to the GDR to achieve these standards, including additional required metadata input fields for geospatial data files and a pipeline to convert geospatial data within the GDR's data lake into the preferred, cloud-optimized, standardized format.

2.1 Geospatial Metadata Standard

According to the International Organization for Standardization (ISO), geospatial metadata attributes should provide comprehensive information that helps users understand, evaluate, and use the data effectively. Some of the key metadata attributes that should be included with geospatial data are:

- **Identification Information**: This includes the title, abstract, purpose, and status of the geospatial data, as well as any keywords that describe its content. Within the GDR, this is already required in the submission form.
- **Data Quality Information**: This includes information on the positional accuracy, attribute accuracy, logical consistency, and completeness of the geospatial data. Within the GDR, this should be included in the submission abstract, resource description(s), or in a readme file.
- **Type of Geospatial Data:** This includes an explanation of the type of geospatial data. For example, is the file geography data (e.g., a csv file with information about coordinates) or geometry data? If the file contains geometry data, is it vector (e.g., points, lines, or polygons) or raster data (e.g., a georeferenced map)? Currently, this information is not required by the GDR submission form, but the GDR team is working on building capabilities to derive this information from file types and adding required inputs for these attributes for ambiguous file types.
- **Coordinate Reference System (CRS)**: This includes information on the coordinate system, units, projection, and datum used for the geospatial data. Currently, this information is not required by the GDR submission form, but the GDR team is working on adding required inputs for these attributes.
- **Name of Geometry Column**: If the geospatial data file contains columnar geometry data, the name of the column containing geometry information should be provided. Currently, this information is not required by the GDR submission form, but the GDR team is working on adding required inputs for these attributes where relevant.
- **Temporal Reference Information**: This includes information on the time period of the data, such as the date of creation, publication, or last update. Within the GDR, the creation date and publication date are accounted for in the GDR submission form, but the time period of the data should be included in the submission abstract, resource description(s), or in a readme file.
- **Data Source Information**: This includes information on the originator, publisher, and any other relevant sources of the geospatial data. Within the GDR, some of this info is already required in the submission form (originator and publisher), but any other relevant sources of the geospatial data should be specified in the submission abstract, resource description(s), or in a readme file.
- **Entity and Attribute Information**: This includes information on the features, attributes, and attribute values present in the geospatial data, as well as any attribute definitions, domains, and units of measure. Within a GDR submission, this information should be specified in the submission abstract, resource description, or an associated readme file.
- **Distribution Information**: This includes information on the format, size, and access methods for the geospatial data, as well as any fees, restrictions, or licensing requirements. The format and size are automatically displayed for every data file in the GDR, but any specific access methods should be specified in the submission abstract, resource description, or an associated readme file.

- **Metadata Reference Information**: This includes information on the metadata itself, such as the date of creation, contact information for the metadata author, and any metadata standards or profiles used. The metadata creation date, contact information, and author(s) are already required within the submission form. Any metadata standards or profiles used which differ from the GDR's standards should be specified in the submission abstract, resource description, or in an associated readme file.
- **Spatial Domain Information**: This includes information on the geographic extent of the geospatial data, such as bounding coordinates or a description of the area covered. This information is required by the location attribute associated with each resource uploaded to a GDR submission.
- **Lineage Information**: This includes information on the history of the geospatial data, such as the methods used for data collection, processing, and quality control. This information should be specified in the submission abstract, resource description, or an associated README file.

By including these metadata attributes with geospatial data, users can better understand the context, quality, and limitations of the data, making it easier for them to use the data effectively in their own analyses and applications (ISO 19115-1).

2.2 Preferred Geospatial Data Formats

The GDR prefers open-source formats with metadata embedded. Some common open formats include:

- **GeoTIFF**: A georeferenced raster image format that is widely used for satellite imagery and aerial photography.
- **Shapefile**: A vector data format developed by Esri, commonly used for storing points, lines, and polygons.
- **NetCDF**: Network Common Data Form (NetCDF) is a set of software libraries and machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data, often used for gridded data and climate models.
- **GeoJSON**: A lightweight format for encoding geographic data structures, often used for web mapping applications, particularly for vector data.
- **GML**: Geography Markup Language (GML) is an XML-based format for encoding geographic information, including both spatial and non-spatial properties of geographic features. GML can be used for both raster and vector data.
- **KML/KMZ**: Keyhole Markup Language (KML) is an XML-based format for storing vector-type geographic data and associated content, often used with Google Earth. KMZ is a compressed version of KML.
- **HDF5**: Hierarchical Data Format, version 5 (HDF5) is a data model, a set of open file formats, and libraries designed, in the context of geospatial data, to store and organize large amounts of raster data for improved speed and efficiency of data processing (The HDF Group, 2023). HDF5 is not natively cloud-optimized but can be through the use of third-party libraries or services (e.g., kerchunk (Durant, 2021) or Highly Scalable Data Service (HSDS, The HDF Group, 2023)).
- **GeoParquet**: GeoParquet is an incubating Open Geospatial Consortium (OGC) standard that adds interoperable geospatial types (Point, Line, Polygon) to Parquet, which is a

column-oriented modern alternative to CSV files (GeoParquet, 2023). GeoParquet is preferred by several cloud service providers for big geospatial data, due to its columnar data format (beneficial to data science workflows), native cloud-optimization, and cloud data warehouse interoperability. *GeoParquet is the GDR's recommended format for large or especially complex vector datasets.*

These formats are widely used and supported by various geospatial software and tools, making them suitable for geospatial data storage and exchange (ISO 19115-1). Of these formats, the GDR prefers GeoTIFF, Shapefile, NetCDF, or GeoJSON, depending on the data's application. The GDR discourages the use of proprietary formats like GeoDatabase because they are not easily accessible to those who do not have licenses for the software required to view and work with them.

High resolution geospatial data are typically made accessible via the GDR data lakes rather than the traditional access model of downloading datasets. Since data lakes make large amounts of data available for use within the cloud, cloud-optimized formats are preferable to improve computational performance, storage cost efficiency, and scalability.

2.3 Geospatial Data Pipeline

To ensure proper metadata requirements for geospatial data, the GDR is being updated to auto-recognize geospatial files and determine what type of geospatial data file it is (i.e., geography versus geometry, vector versus raster). The resource-specific metadata will have additional required input fields to specify a single CRS for each file, along with the name of the geometry column if relevant. The CRS field accounts for the coordinate system, units, projection, and datum used for the geospatial data. The name of the geometry column enforces the inclusion of a single CRS for each vector geospatial data file stored in a columnar format.

In addition, a pipeline to convert geospatial data into GeoParquet format is being developed as part of the Open Energy Data Initiative (OEDI), an effort to improve access to valuable datasets and automate data management across the Department of Energy's (DOE) programs. This data pipeline will be integrated into the GDR data lakes for large, high resolution geospatial data.

3. DAS Data Standards

Since there are already existing data and metadata standards for DAS data (i.e., PRODML Work Group, 2022 and IRIS DAS RCN Data Management Working Group, 2022), the GDR is building off of these and focusing its efforts on the pipeline to achieve these standards.

3.1 DAS Metadata Standard

For a DAS metadata standard, the GDR suggests following the recommendations of the IRIS DAS RCN Data Management Working Group. The IRIS DAS RCN Data Management Working Group breaks the metadata requirements into five major blocks: overview, cable and fiber, interrogator, acquisition, and channel. Each block includes a list of both required and optional fields, along with a definition, type, format, additional instructions, and examples for each attribute. Here we provide a summary of the metadata guide, but the most complete and up-to-date information can be found

on the IRIS DAS RCN Data Management Working Group's GitHub repository[1]. A glossary of these terms along with additional background information may be found in the IRIS DAS RCN Data Management Working Group Whitepaper.

Overview metadata provides high-level information about the DAS deployment and helps to facilitate discovery based on spatiotemporal searches. It includes location, deployment type, network, site name, number of interrogators, principal investigators, start datetime, and end datetime as required fields.

Cable and fiber metadata describes the cable environment and the fibers used within the cable(s) used over the course of an experiment. This metadata aims to uniquely specify the fiber used to collect measurements. It includes cable fiber ID and cable coordinates as required fields.

Interrogator metadata provides information about the interrogator(s) used to collect the data during an experiment. Each interrogator gets its own metadata block, each including a unique identifier, the manufacturer and model of the interrogator, and the units of measure.

Acquisition metadata contains information about data collection parameters and signal processing steps. It requires a user-defined acquisition ID, an acquisition start time, acquisition end time, acquisition sample rate, gauge length, number of channels, channel spacing, archived sample rate, units of measure, decimation applied to the data, and filtering process(es) applied to the data.

Channel metadata describes each individual channel along the fiber. Like with the interrogator, each channel gets its own metadata block. It requires a name of the associated file, file format of associated file, generation date, channel. ID, reference frame, location method, and direction of laser pulse. It also requires a coordinate file with the channel ID, distance along fiber (km), X and Y-coordinates, and depth (km).

3.2 Preferred DAS Data Format

The GDR models its standard format after PRODML v2.2 (PRODML Work Group (Energistics), 2022). Industry feedback tells us that it is challenging to force providers of DAS interrogators or services to record data in a particular format. However, over the last two years, there has been progress towards standardization, as most of the major DAS vendors have added a PRODML export option, in addition to their proprietary formats, meaning that researchers involved in collecting DAS data can request the data in PRODML format for their projects. Asking them to add another format would be a challenging feat, so using PRODML to model the GDR's data standard is the most logical and synergistic approach. Additional collaboration between Energistics, the PRODML Work Group, the IRIS DAS RCN Data Management Working Group, and the GDR Team may be needed to synchronize the above DAS metadata standard with PRODML.

Within PRODML, XML files are used for storing DAS metadata, due to their machine-readability and human-readability. Hierarchical Data Format, version 5 (HDF5) is used to store raw and processed DAS data. HDF5 is a data model, a set of open file formats, and libraries specifically designed to store and organize large amounts of numerical/array data for improved speed and

[1] https://github.com/DAS-RCN/DAS_metadata/blob/main/term/README.md

efficiency of data processing (The HDF Group, 2023). Since both the raw and processed datasets can be useful, The GDR recommends hosting raw data beside the processed data. The data stored in HDF5 format consists of both raw and processed arrays. The HDF5 file contains necessary ancillary and metadata attributes for the groups and arrays. The file structure and naming conventions must follow the specified guidelines. The metadata is duplicated in both the XML and HDF5 files to ensure coherence in case the files become physically separated during transit. DAS data arrays can be very large, and it is possible to split arrays across multiple physical HDF5 files (PRODML Work Group, 2022). Within the GDR's data standard, the GDR considers separate metadata XML file as optional, since the metadata should be stored directly in the data files and is also included in the GDR submission form for human-readability.

HDF5 files are comprised of groups to organize data elements, datasets (i.e., arrays), to store actual data, and attributes to provide metadata. It supports various data types, data links, and flexible storage approaches, making it a versatile format for managing complex datasets in scientific and engineering applications (The HDF Group, 2023). In each HDF5 file, the following groups are recommended to be stored:

1. '**DasMetadata'**: A group of all the DAS required metadata attributes, as described in Section 3.1, and their associated values.
2. '**DasRawData'**: A group including a data array of raw DAS data ('RawDataArray') along with the dimensions of the data ('DasDimensions').
3. '**DasSpectraData'**: *If applicable.* A group including a data array of Fourier transformed spectrum data ('SpectraDataArray'), 'StartFrequency,' 'EndFrequency,' and the dimensions of the data ('DasDimensions').
4. '**DasFbeData'**: *If applicable.* A group including a data array of frequency band extracted (FBE) data ('FbeDataArray'), 'StartFrequency,' 'EndFrequency,' and the dimensions of the data ('DasDimensions').
5. **'Das[*OtherProcessingTechnique*]Data'**: If other processed forms of the DAS data exist, they should be included as additional groups with appropriately named data arrays. This array should also include any other relevant processing parameters, named intuitively, and the dimensions of the data ('DasDimensions').
6. **'DasTimeArray'**: A group including a datetime index array for DAS dataset ('TimeArray'), start TimeStamp ('StartTime'), and end TimeStamp ('EndTime').

Within PRODML, Energistics Packaging Conventions (EPC) format is useful for grouping multiple files together as a single package (or file), which makes it easier to exchange the many files that may make up a data model. EPC is an implementation of the Open Packaging Conventions (OPC), a commonly used container file technology standard supported by two international standards organizations. Essentially, an EPC file is a .zip file that can be opened and viewed using any .zip tool (PRODML Work Group, 2022). Within the GDR's data standard, use of the EPC format is not required. Instead, it is recommended to upload the HDF5 files to the GDR's data lakes, rather than through the traditional upload model, to enable working with the data directly in the cloud. HDF5 can be cloud-optimized using third party services such as kerchunk (Durant, 2021) or Highly Scalable Data Service (HSDS, The HDF Group, 2023) to make it more convenient, scalable, and cost- and performance-efficient to use in cloud environments.

3.3 DAS Data Pipeline

The GDR is currently developing a pipeline to automatically convert DAS data submitted in SEG-Y format into the GDR standard HDF5 format.

4. Impact and Benefits

The implementation of the automated data pipelines discussed in this paper will significantly enhance data interoperability and integration, facilitating efficient access and analysis of big geospatial and DAS datasets from various sources. By incorporating standardized metadata requirements for these datasets, data discovery and usability will be further improved. Researchers will be able to better access detailed information about geospatial datasets and DAS recordings, such as CRS, location, acquisition parameters, and sensor characteristics, through consistent metadata. This standardized approach would streamline the use of data uploaded to the GDR from different sources, saving valuable time.

Moreover, within data-centric approaches to machine learning and artificial intelligence, high quality input data is a key aspect of achieving high quality machine learning results. One aspect of high-quality data is structure and standardization. By applying standardization techniques to big geospatial and DAS data through automated data pipelines, researchers will ultimately see higher quality outputs from their DAS or geospatial machine learning projects. Data standardization would act as a step zero in the data curation process described by Taverna et al. (2023), easing data digestion and transformation, and leading to a more efficient production of high-quality machine learning outputs. The use of standardized datasets would also enable researchers in the geothermal community to explore a wider range of machine learning experiments and interpretations, fostering more applicable outcomes for real-world geothermal challenges.

Automated data pipelines, along with standardized data formats and metadata requirements, offer several benefits, including improved data quality, consistency, and collaboration. Leveraging data lakes as central, cloud-based data stores, researchers can access and share big geospatial and DAS datasets seamlessly, enabling real-time collaboration among geographically dispersed teams. This approach not only accelerates analysis but also promotes a culture of collaboration, leading to advancements in scientific knowledge within the geospatial and geothermal research fields.

5. Challenges

One of the main challenges associated with implementing these data standards and pipelines is addressing data format and compatibility issues, including metadata formats. For example, there are slight variations in formats like SEG-Y, and ad-hoc formats for many other datasets (e.g., stimulation data). This makes it challenging to develop comprehensively compatible data pipelines. To circumvent this challenge, the GDR focuses on the most commonly used formats, and utilize flexible, modular pipeline development to ease integration of slight variations from common formats.

Another challenge is ensuring data and metadata completeness and accuracy for metadata attributes, particularly those that get incorporated into more open-ended fields in the submission form, such as the submission abstract, resource descriptions, or an optional readme file. Since the submission form is intended to be flexible to any dataset being submitted, the GDR cannot mandate inclusion of the same metadata in these fields for every submission, which means that the burden of ensuring complete metadata can be somewhat ad-hoc and falls on the curation team. While the

curation team is strong, it is comprised of humans, making them susceptible to mistakes occasionally. The GDR is combatting this challenge through thorough training of the curation team on diverse data types, attempting to standardize the curation process through the use of a checklist, and through improved documentation of best practices for submitting data and associated metadata to the GDR.

Another challenge is related to the costs associated with storing such large datasets, and the compute costs associated with operating the big data pipelines. OEDI, in collaboration with major cloud providers such as Amazon Web Services, Google Cloud, and Microsoft Azure, has developed cloud-based data lakes that store over 1 petabyte (1 PB) of publicly accessible data, with storage costs covered by the cloud providers' public data programs. The GDR team does, however, still need to cover the cost of data translation using project funds. These costs are currently covered by OEDI and the GDR, but future big data pipeline throughput could potentially be limited by budget.

Lastly, standards and preferred formats are constantly evolving (i.e., new and further cloud-optimized formats, new and improved versions of existing formats). This makes it challenging for the GDR data standards and pipelines to stay current with the state-of-the-art. This problem is mitigated through regularly scheduled maintenance and improvements to the data standards and pipelines centered around updates and shifting paradigms.

6. Conclusion

The integration of automated data pipelines and data standardization is crucial to advancing geothermal machine learning research. Such an approach enhances data accessibility, quality, and collaboration, fostering more effective and applicable machine learning outcomes for addressing real-world challenges in geothermal energy and other domains.

To help achieve this, the GDR is implementing data standards and pipelines for high value datasets, including drilling data, DAS data, geospatial data, and potentially stimulation data in the future. Following the implementation of these data standards and pipelines, the GDR team is planning on making refinements to the existing standards and pipelines. If these standards and pipelines continue to provide value, more will be developed and implemented.

Lastly, the GDR team is continuously working to align its efforts with the needs of the geothermal community. That said, the GDR team would like to invite you to provide your feedback on the existing standards and pipelines, or suggestions for future data standards and pipelines, here: GDRHelp@ee.doe.gov.

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 United States Government. The United States Government retains and the publisher, by accepting the

article for publication, acknowledges that the United States 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 United States Government purposes.

We'd also like to acknowledge Aleksei Titov at Fervo Energy for his review and input related to geothermal industry adoption of a DAS data standard, and OpenAI's ChatGPT 3.5 and 4 for its assistance in summarizing and distilling technical usage guides for existing geospatial metadata (ISO) and DAS data (PRODML) standards. In addition, we'd like to acknowledge Meghan Mooney at NREL for her review of and suggestions around the geospatial metadata and data standards proposed in this paper, and Jianli Gu at NREL for his work to develop a pipeline to convert geospatial data into GeoParquet format as part of the Open Energy Data Initiative (OEDI).

Lastly, we'd like to thank the GDR curation team, currently Nathan Danigelis, Scott Mello, and Adrienne Lowney, for their continued dedication to ensuring high-quality metadata associated with submissions.

REFERENCES

The Geothermal Data Repository (GDR). https://gdr.openei.org/home.

Durant, M. "kerchunk" (2021). https://fsspec.github.io/kerchunk/.

GeoParquet (2023). https://geoparquet.org/.

Infostat. "RIMBase" (2023). https://infostatsystems.com/well-operating-companies/.

IRIS DAS Research Coordination Network (RCN) Data Management Working Group. "Distributed Acoustic Sensing (DAS) Metadata Model." Whitepaper (2022). https://github.com/DAS-RCN/DAS_metadata.

ISO 19115-1. "Geographic information — Metadata." (2020). https://committee.iso.org/sites/tc211/home/projects/projects---complete-list/iso-19115-1.html.

Nabors Industries Ltd. "RigCloud" (2021). https://rigcloud.com/.

National Geothermal Data System (NGDS). "Data Exchange Models" (2013). https://www.geothermaldata.org/content-models/data-interchange-content-models.

Pason Systems Corp. Pason Systems (2023). https://www.pason.com/.

PRODML Work Group, "PRODML Technical Reference Guide." Energistics, v2.2 (2022). https://www.energistics.org/prodml-data-standards.

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).

Taverna, N., Weers, J., Huugins, 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).

The HDF Group (2023). https://www.hdfgroup.org/.

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).