PROCEEDINGS, 48th Workshop on Geothermal Reservoir Engineering
Stanford University, Stanford, California, February 6-8, 2023
SGP-TR-224

Improving the Quality of Geothermal Data Through Data Standards and Pipelines Within the Geothermal Data Repository

Nicole Taverna[1], Jon Weers[1], Jay Huggins[1], Sean Porse[2], Arlene Anderson[2], Zach Frone[2], and RJ Scavo[3]

[1] National Renewable Energy Laboratory (NREL), Golden, CO 80401, USA

[2] U.S. Department of Energy (DOE), Washington, D.C. 20004, USA

[3] Formerly National Renewable Energy Laboratory (NREL), Golden, CO 80401, USA

Nicole.Taverna@nrel.gov

Keywords: GDR, data, standardization, pipelines, data science, machine learning

ABSTRACT

For machine learning outputs to be applicable to real-world problems, high-quality data are needed to ensure high-quality results. With recent emphasis on machine learning in geothermal, there is an increasing need for greater focus on the quality of the data available for use in these projects. High-quality datasets result from dependable sensors or devices collecting data, high frequency of measurements, sufficient data points, adequate metadata, reliable storage of data, and sufficient data curation. Another component that contributes to high-quality data is reusability, which can be enhanced through data standardization. Data standardization creates consistency in formatting and contents of like datasets, lessening preprocessing requirements and ensuring adequate information provided by a given dataset. The Geothermal Data Repository (GDR)—which houses data from research funded by the U.S. Department of Energy Geothermal Technologies Office—aims to help improve data quality through automated data standardization for high-value datasets through the implementation of data pipelines alongside reliable and accessible long-term storage for datasets. As such, the GDR has decided to shift away from recommending the use of Excel-based content models and toward the implementation of automated data pipelines. This takes the burden of data standardization off the user and project team and will increase the availability of standardized geothermal data available through the GDR. A set of recommendations, or a data standard for each data type, will exist with each data pipeline in order to advise data collection for maximum usability for future research. This paper describes the GDR's proposed transition toward data standardization through automated data pipelines, discusses the need for and value of such a shift, and calls for suggestions from the community regarding the most useful data standards and pipelines.

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 of increasing value to geothermal machine learning projects as this field of research grows in popularity. That considered, the GDR is aiming to improve the convenience and efficiency of using its datasets in geothermal machine learning projects.

1.1 The Importance of High-Quality Data in Machine Learning

For machine learning outputs to be applicable to real-world problems, high-quality data are needed to ensure high-quality results. With recent emphasis on machine learning in geothermal, there is an increasing need for greater focus on the quality of the data available for use in these projects. An example of high-quality data leading to successful project outcomes is the Geothermal Operational Optimization Using Machine Learning (GOOML) project, which utilized large quantities of geothermal power plant operational data to inform power plant operational configurations to maximize power generation (Buster et al., 2021). One of the key components to the success of the GOOML project was the focus on data curation.

The GOOML project aimed, in part, to serve as an example of best practices for data curation within machine learning projects. The process is graphically shown in Figure 1, and consists of the following steps: 1) acquisition of data from data owners, 2) digestion of data to gain an initial understanding of what is included, 3) data transformation, which includes converting the data into a standardized machine-readable format, 4) quality assurance and quality control, involving identification of significant data gaps and apparent anomalies, 5) use in machine learning algorithms, and 6) repetition of steps one through five until all data needs are met and data are deemed suitable for producing trustworthy modeling results that may be disseminated, ideally along with the curated dataset. This process is often iterative, wherein the focus is placed on improving the quality of the data rather than tuning machine learning model parameters. The GOOML data curation process supports a data-centric philosophy, rather than the alternative model-centric approach, with the goal of improving real-world applicability of geothermal machine learning projects (Taverna et al., 2022). Outcomes from the GOOML project along with other geothermal machine learning projects demonstrate that high-quality data are critical for the success of these projects.

1.2 What Constitutes High-Quality Data?

High-quality datasets result from dependable sensors or devices collecting data, high frequency of measurements, sufficient data points, adequate metadata, reliable storage of data, and adequate data curation. Another component that contributes to high-quality data is reusability, which can be enhanced through standardization. Data standardization creates consistency in formatting and contents of like datasets, lessening preprocessing requirements and ensuring adequate information provided by a given dataset.

Preferred data formats for GDR submissions are those that support the best reusability; however, the GDR accepts a variety of file formats and will, in most cases, accept submissions in whatever format they are provided in. For data available in multiple formats, the GDR provides the guideline in Figure 2, which describes the different levels of reusability of data using tiers based on format, in order of increasing inherent reusability. The tiers include: 1) unstructured data, which includes data in PDF, PowerPoint, image, or similar formats. Unstructured data is acceptable, but a better option when possible is 2) structured data. Structured data includes data in Excel, CSV, XML, and other similar formats. The best option, however, is 3) structured and standardized data. This tier includes standardized data in Excel, CSV, XML, RDF, JSON, or other similar formats (e.g., the National Geothermal Data System content models, discussed next).

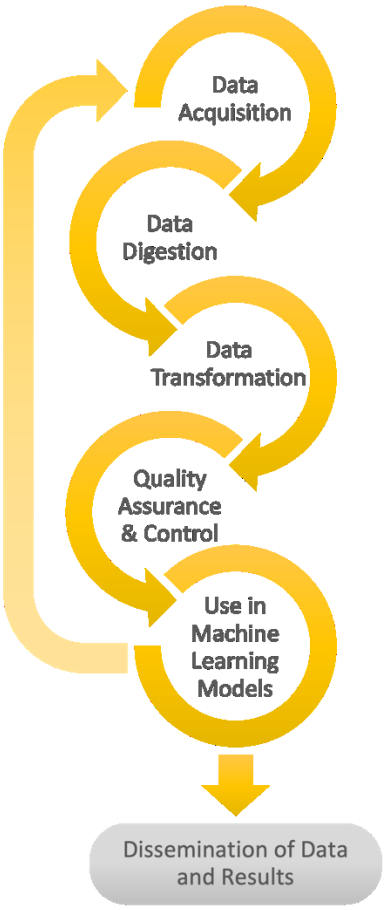

Figure 1: Graphic describing the data curation process used by the GOOML project.

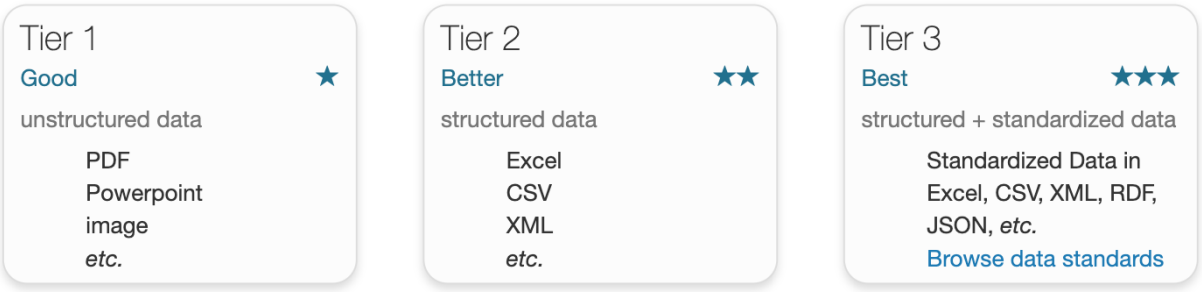

Figure 2: Graphic describing the GDR's guideline for preferred data formats. In this guideline, Tier 3: structured + standardized data is considered best because it maximizes reusability.

1.3 GDR's Previous Approach to Data Standardization

In the past, users were encouraged to use the National Geothermal Data System (NGDS) content models as a means of standardizing their data. NGDS provides standardized templates in Excel and XML formats for users to input their data. The NGDS content models were

developed with the intent of being all-inclusive, meaning that there is a column for every possible measurement associated with a particular data type. Table 1 shows the data types that NGDS content models were developed for (NGDS, 2013).

While simple to use and understand, the NGDS content models put the burden of data standardization on the submitter. This is often problematic because data submission usually happens at the end of a project, when project funds may be running low. This makes submitters less inclined to add the extra step of manual data standardization—especially when it is optional. Additionally, by requiring data to be stored in either Excel or XML format, the NGDS content models are limited to capturing data compatible with these formats (Weers et al., 2021). Their ability to support time-series data or big data such as Continuous Active-Source Seismic Monitoring (CASSM) or Distributed Acoustic Sensing (DAS) data is limited at best and they offer no support for the standardization of non-tabular data such as core photos, video files, or complex geospatial data.

Table 1: List of NGDS content models available for use.

Abandoned Mines	Active Fault / Quaternary Fault	Aqueous Chemistry
Borehole Lithology Intercepts	Borehole Lithology Interval Feature	Borehole Temperature Observation
Contour Lines	Direct Use Feature	Drill Stem Test Observations (deprecated)
Fault Feature / Shear Displacement Structure	Fluid Flux Injection and Disposal	Geologic Contact Feature
Geologic Fault Feature / Shear Displacement Structure	Geologic Reservoir	Geologic Units
Geothermal Area	Geothermal Fluid Production (deprecated)	Geothermal Metadata Compilation
Geothermal Power Plant Facility	Gravity Stations	Heat Flow
Heat Pump Facility	Hydraulic Properties	Mineral Recovery Brines
Physical Sample	Powell and Cumming Geothermometry	Power Plant Production
Radiogenic Heat Production	Rock Chemistry	Seismic Event Hypocenter
Thermal Conductivity Observation	Thermal/Hot Spring Feature	Volcanic Vents
Well Fluid Production	Well Header Observation	Well Log Observation
Well Tests		

2. APPROACH

2.1 NGDS Content Model GDR Upload and Download Statistics

Considering the limitations of the NGDS content models, it is not surprising that they suffer from low adoption, as is evident by GDR metrics (see Table 2). There are 39 total content models in GDR (0.0073% of all resources). The top 3 are Aqueous Chemistry (20), Mineral Recovery Brines (8), and Geologic Reservoirs (3), although the more frequently used content models are more indicative of funding opportunities and data management requirements than they are indicative of demand. To get a better idea of demand, we looked at downloads from the GDR. Since 2019, there were 2,994 total downloads of content models, with the top three downloaded content model types being: Aqueous Chemistry (1151), Well Log Observations (680), and Mineral Recovery Brines (515). To remove the bias associated with there being more of some content models in the GDR than others, we also calculated the top three average downloads per content model type, with the top three being: Well Log Observations (680), Geologic Reservoirs (84), and Heat Flow Features (81). These download statistics suggest that there is one somewhat popular Well Log Observations content model, but that otherwise, the content models are not in high demand.

Table 2: GDR NGDS content models upload and download statistics.

Content Model	Number of Instances within the GDR	Downloads (total for all content models of specified type in submission)	Downloads (averaged over all content models of specified type in submission)
Aqueous Chemistry	20	1151	58
Borehole Temperature Observation	1	58	58
Direct Use Features	1	67	67
Geologic Reservoirs	3	253	84
Heat Flow Features	1	81	81
Hydraulic Properties Observations	1	37	37

Mineral Recovery Brines	8	515	64
Rock Chemistry	2	111	56
Seismic Event Hypocenter Observations	1	41	41
Well Log Observation	1	680	680
All	39	2994	77

2.2 Automated Data Pipelines and Standards

Based on the NGDS Content Model upload and download statistics, the GDR has decided to shift away from recommending the use of Excel-based content models and toward the implementation of automated data pipelines. 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 3). This shift takes 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. A set of recommendations and a data standard for each data type will exist with each data pipeline in order to advise data collection for maximum usability for future research.

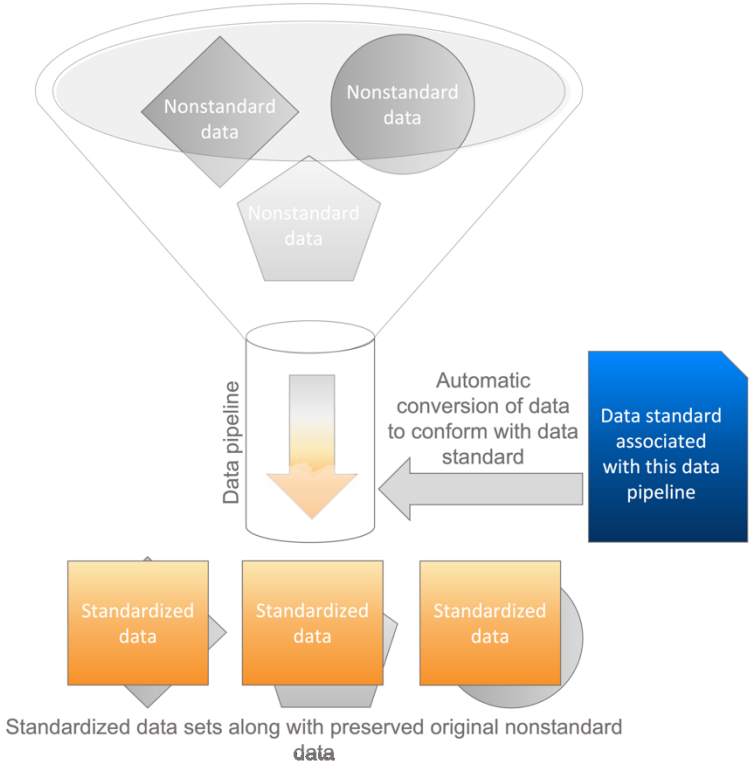

Figure 3: Graphic depicting how GDR data pipelines work. Nonstandard data is uploaded to the GDR and is funneled through a data pipeline that standardizes the data in accordance with the associated data standard, while preserving the original nonstandard data.

At the time of writing this paper, the GDR has implemented an automated data pipeline and data standard 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 and can easily be amended to standardize drilling data from other sources as well. The standard, of which the first few data fields are shown, is displayed in Figure 4, and additionally includes RIMBase-specific (Infostat, 2023) field names. It provides a table of standard data fields, their platform-specific fields names, the appropriate standard units, and whether the field is required or not. It also serves as a map for how the pipeline works. It recognizes the platform-specific field names and units and converts them to the standard field names and units.

See all standards

Drilling Data Standard

Field	Drilling Data Field Names	Standard Units	Required
AutoDriller Block Velocity *A measure of how fast the traveling block is moving up or down the hole.*	Pason Field Name: **AutoDriller Block Velocity** RIMBase Field name: RigCloud Field Name:	feet per hour (ft/hr)	
Azimuth *The compass direction of the wellbore as measured while drilling. Usually specified in degrees with respect to the geographic or magnetic north pole.* *Include whichever is available.*	Pason Field Name: **Azimuth** RIMBase Field name: **Azimuth In** RigCloud Field Name: **MWD Azimuth**	degrees	Required

Figure 4: Subset of the GDR drilling data standard. The full drilling data standard may be found here: https://gdr.openei.org/standards/drilling_data.

When a submitter uploads a drilling dataset to the GDR, the process is similar to uploading any other data file. The submitter will need to add resource-specific metadata, including a display name, file description, date, and location. The process does not differ until the submitter hits "Save" or "Submit." After saving or submitting the submission, the data pipeline detects the drilling data file and its native format and begins converting the data into the data standard, automatically generating an additional file that appears in the resource list (Figure 5). The file will initially appear as grayed out with a tool tip (i.e., the gray circle with an 'i' in it). If the submitter hovers over the tool tip, they will see that the file is an auto-generated one by the drilling data pipeline. If they click on the tool tip, it will take them to the drilling data standard page (Figure 4). The file will remain grayed out and not available for download until the standardization is complete, which typically only takes a few minutes or less, depending on the size of file. Then it will appear as a normal resource, but the tool tip will remain. The resource-specific metadata are auto-filled metadata from the original file but may be edited if needed.

After drilling data are submitted and standardized, they will appear to the general user as an additional file available for download (Figure 6). This allows the original data to be preserved for posterity and be presented alongside the standardized data. Note the overlap and distinctions in name and description between the original file and the standardized version.

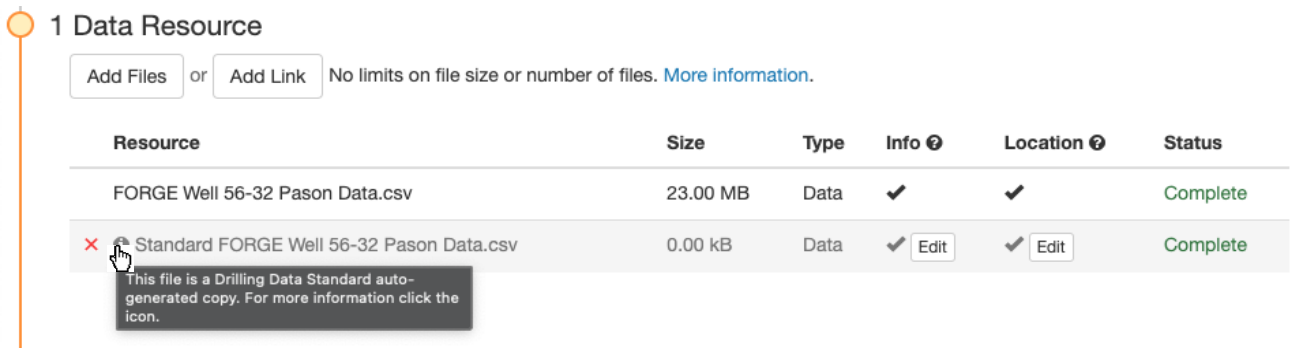

Figure 5: Screenshot of what the user sees after uploading a drilling data file, filling out the resource-specific info, and hitting save.

Figure 6: Screenshot showing the general user's view of a drilling data file and its standardized counterpart. Note the overlap and distinctions in name and description between the original file and the standardized version.

2.3 Determining Priorities for New Data Standards and Pipelines

In selecting priorities for new data standards and pipelines, the GDR team wanted to ensure alignment with anticipated awards, high-demand datasets, and feedback from the community. To do this, the team began with internal brain-storming sessions, and then conducted an internal survey (of staff at the National Renewable Energy Laboratory) on what the geothermal team would most like to see improvement to in the GDR. This survey yielded feedback mostly related to geospatial data, wherein datasets often do not include a coordinate reference system (CRS), leaving the burden on the user to try to determine the CRS for a dataset in question—a task that can be extremely time consuming and sometimes impossible to determine with confidence.

The GDR team then hosted discussions with members of the DOE GTO discussing data-related priorities for FY23 to ensure our efforts and ideas aligned. Particularly, we wanted to ensure that we were accounting for specific types of existing datasets that may become popular due to upcoming work, or common outputs from the work planned for FY23 and soon after, especially related to high-profile projects such as FORGE, EGS Collab, and heat pump and low-temperature research planned for FY23.

Following these discussions and analysis, the GDR team is proposing the following new data pipelines and data standards for FY23:

1. Geospatial datasets: The GDR will auto-detect geographic information system (GIS) data files using their extensions and require users to include complete metadata for GIS data files. This GIS-specific metadata will not only improve reusability of GDR GIS data, but will also enable future integration of GDR GIS data within the proposed Geothermal Energy Atlas.

2. DAS data: The GDR will develop a data pipeline that detects non-standard DAS data and converts it into a standard format in line with the IRIS DAS Research Coordination Network Data Management Working Group's soon-to-be-finalized DAS Metadata Model (Mellors et al., 2022).

3. Stimulation data: If time and funding permit, a third data standard and pipeline will be developed for stimulation data. This is to include injection flow rate, pressure, temperature, and time. This type of data would be auto-detected and converted into a standardized csv format.

3. DISCUSSION

High-quality data is a key component for producing high-quality machine learning results and for the applicability of machine learning to real-world problems. Machine learning is frequently exploratory in nature, meaning that data curation is often an iterative process throughout the life of a project. Modular data pipelines and standardization of processes and practices help to streamline this process, supporting the move to data-centric machine learning workflows that often produce outcomes that are more applicable to real-world problems (Taverna et al., 2022).

That said, any processes that can be taken to lessen data curation requirements are helpful in improving geothermal machine learning results. Data standardization puts similar datasets into a standard format, which lessens the time spent by researchers reformatting and combining datasets. If data standardization were to be added into the GOOML data curation process (Figure 1), it would act as a step zero, which would ease the data digestion and data transformation steps, and likely reduce the number of iterations required to produce high-quality machine learning outputs. This would overall reduce the amount of time required for adequate data curation, both reducing the overall cost of machine learning projects and allowing more time for exploring different machine learning experiments and properly interpreting results. It would also allow users to incorporate many datasets efficiently and possibly automatically into machine learning projects, as opposed to focusing on just one or a few manually parsed datasets.

By taking the burden off the submitter and automating the standardization process for high-value datasets, the availability and hopefully the use of standardized datasets will increase. This would mean that the benefits of data standardization would be felt more widely by the geothermal community, leading to better overall outcomes from geothermal machine learning projects.

4. CONCLUSIONS AND FUTURE DIRECTIONS

The GDR is lessening the data curation requirements associated with geothermal machine learning projects through implementing automated data pipelines to standardize high-value datasets. Providing access to structured and standardized data eases data digestion and data transformation and reduces the number of iterations required to produce high-quality machine learning outputs. This fiscal year, the GDR is planning to prioritize implementing data standards and pipelines for geospatial, DAS, and stimulation data. If the data standards and pipelines continue to provide value to the geothermal community, more will be developed in the future.

Lastly, the GDR team is continuously working to align its efforts with the needs of the geothermal community, and we would like to invite you to provide your feedback here: GDRHelp@ee.doe.gov.

5. ACKNOWLEDGEMENT

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. 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.

REFERENCES

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

Buster, G., Siratovich, P., Taverna, N., Rossol, M., Weers, J., Blair, A., Huggins, J., Siega, C., Mannington, W., Urgel, A., Cen, J., Quinao, J., Watt, R., and Akerley, J. A New Modeling Framework for Geothermal Operational Optimization with Machine Learning (GOOML). Energies (2021), 14, 6852.

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

Mellors, R., Hodgkinson, K., Hui Lai, V., and the DAS Research Coordination Network Data Management Working Group. Distributed Acoustic Sensing (DAS) Metadata Model. Whitepaper (2022). https://github.com/DAS-RCN/DAS_metadata.

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/.

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 Geothermal Program (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).

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).