Improving the Accessibility and Usability of Geothermal Information with Data Lakes and Data Pipelines on the Geothermal Data Repository

Jon Weers [(a)], Sean Porse [(b)], Jay Huggins [(a)], Michael Rossol [(a)], and Nicole Taverna [(a)]

[(a)] National Renewable Energy Laboratory (NREL)

15013 Denver West Parkway

Golden, CO 80401-3305

[(b)] U.S. Department of Energy (DOE)

1000 Independence Ave. SW

Washington D.C. 20004, USA

Keywords

Geothermal, data, lake, repository, GDR, OpenEI, OEDI, seismic, FORGE, PoroTomo, drilling, DAS, DTS, DSS, DOE, collaboration, big, storage, transfer, dissemination, access, open, discoverability, usability, accessibility, standards, pipeline, translation, detection, equality

ABSTRACT

The Geothermal Data Repository (GDR) provides universal access to data and information resulting from research and development activities funded by the Department of Energy (DOE). The GDR has extended this universal access to big data through integration with data lakes developed by the Open Energy Data Initiative (OEDI). Previously, large datasets such as seismic waveform or distributed acoustic sensing (DAS) data could only be accessed by institutions with high performance data storage and compute capabilities, effectively limiting the accessibility of big data to national labs, larger universities, and major corporations. Moreover, the time and resources needed to transport big data and configure them can produce additional barriers to use. Many of the standard formats used for structured data models (also known as content models) are incapable of handling big data and can introduce additional usability problems, often requiring data to be reformatted prior to use. This paper will explore how recent integrations between the GDR and the OEDI data lake have improved the accessibility and usability of geothermal data in a big way, making the data available to a broader audience, and enabling collaborative analysis and innovation across the greater geothermal industry.

1. Introduction

The volume of data being generated by geothermal research, development, and operations is increasing exponentially. The number of activities generating large amounts of data is increasing while the data being generated are also increasing in both size and complexity, due in part to advances in sensing technology allowing for more sensors to be deployed and at higher resolutions. Our ability to generate data is exceeding our capacity to efficiently store, manage, and access those data (Weers and Anderson 2016). Classic paradigms for data management and access are struggling to keep up with "big data" formats, such as distributed acoustic sensing (DAS) data. Additionally, these classic architectures can create barriers to data access and collaboration.

2. Barriers to Access

Data and information sharing platforms such as the U.S Department of Energy's (DOE's) Geothermal Data Repository strive to eliminate barriers to data access in order to disseminate their contents to the broadest scientific community possible. Numerous intentional architectural design decisions and strategic development efforts have been made to improve the discoverability of data within these systems and help ensure that their contents are universally accessible (Weers and Anderson 2015). This helps to ensure that data stored within the catalog is accessible by anonymous users, without requirements for registration, payment, institutional access limits, or other socioeconomic barriers.

2.1 Challenges of Big Data

The challenges associated with transferring, storing, and utilizing big data can present their own barriers to access. Traditionally, big data could only be stored in large arrays of physical drives and often required high performance compute (HPC) capabilities in order to use them. The physical space requirements and costs associated with maintaining such capabilities introduced technical and financial barriers to the accessibility of big data, often restricting access to national labs, larger universities, and major corporations.

Collaboration on big data was limited to only those partnering organizations who also possessed similar capabilities, or to whom institutional access had been granted at one of the capable partnering organizations. This would often introduce institutional barriers to data access, resulting from restrictions on partner access eligibility and incompatibility with internal data organization schemes or HPC environment configurations.

Simply sharing these data with potential collaborators presents its own challenge, requiring specialized peer-to-peer data transfer tools (e.g., Globus) that must be installed by both parties and configured to be operated through institutional barriers. Each individual data transaction had to be coordinated and monitored by both parties to ensure successful completion, limiting the accessibility of those data to the availability of the transferring organization.

A lack of available guidelines and standards for big data formats presents another challenge. Current industry standards for the organization of geothermal data leverage inefficient templates and rely on structured data models limited to tabular data. Also known as "content models", these data models often require data producers or submitters to translate their data into spreadsheet-based templates or text-heavy markup languages such as XML. These are often not

practical for use with big datasets, which can span millions or even billions of rows. Popular spreadsheet editing programs have limits to the number of rows they support (Microsoft 2021), and the bloat incurred from wrapping values in a markup language can double or even triple the size of a dataset, making it considerably more difficult to access and use.

3 Overcoming Big Data Challenges

The DOE Geothermal Data Repository (GDR) is adopting several modern data architectures and workflows to address many of the challenges of working with big data, including cloud-based data lakes, modern standards, and data pipelines. At its core, the GDR is a data repository and information dissemination tool designed to protect DOE's investment in research and development activities by preserving data from those activities and communicating the resulting information to the greatest audience possible. The metadata for all datasets submitted to the GDR are federated (i.e., transmitted) to a network of data sharing partners (Figure 1), greatly increasing the exposure and discoverability of GDR data.

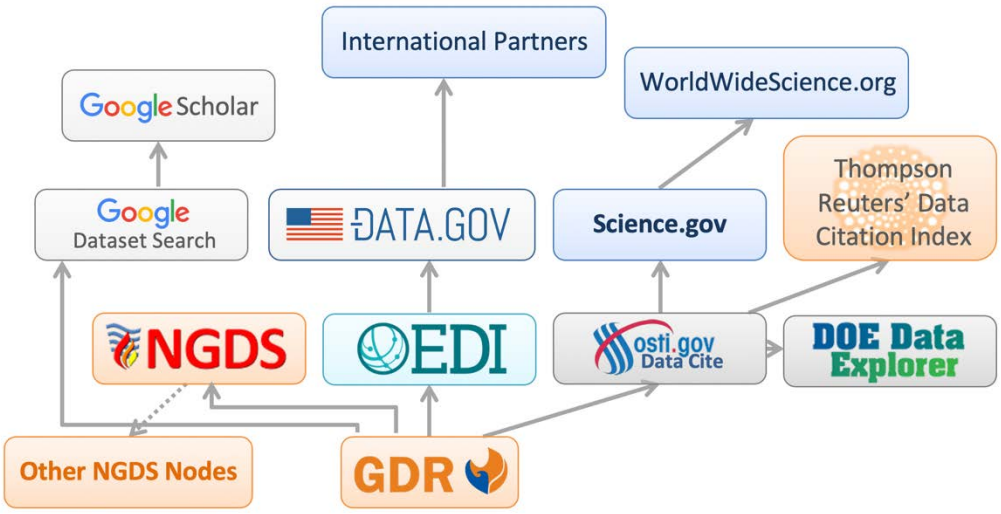

Figure 1. Diagram of GDR metadata propagation through a network of data sharing partners.

In addition to making DOE geothermal data discoverable through the GDR application itself, the data are also discoverable through each of the platforms listed above, without any duplication of effort or redundant storage. Users of these platforms can search, discover, find, and download GDR data seamlessly from each platform, just as they would any other dataset, while behind the scenes the download request is being routed back to the GDR. This exposure to a broader user base increases the discoverability of GDR data by a factor of 12, based on analytics captured by GDR servers that compare direct downloads to those originating from the network of data sharing partners.

Universal accessibility is an important part of the GDRs data management and dissemination strategy. In order to support big data submissions and overcome potential barriers introduced by larger datasets, the GDR has integrated with DOE Open Energy Data Initiative (OEDI) to house

select datasets in OEDI's cloud-based data lake and support modern research paradigms. Furthermore, new data standards have been developed for the GDR and integrated into its catalog using data pipelines to better facilitate data transfer and overcome some of the barriers of standardizing big data.

4 Data Lakes and the Open Energy Data Initiative (OEDI)

The Open Energy Data Initiative (OEDI) improves and automates access to high-value datasets across DOE's programs. Developed and maintained by the National Renewable Energy Laboratory (NREL), OEDI makes data actionable and discoverable by researchers and industry to accelerate analysis and advance innovation through the development of several cloud-based data lakes, in collaboration with major cloud providers, including Amazon Web Services, Google Cloud, and Microsoft's Azure. OEDI is already home to more than 1 petabyte (1 PB) of widely accessible, public data.

Data lakes are modern data storage and access architectures that represent a departure from traditional research paradigms. 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 numerous inefficiencies. In the data lake architecture (Figure 2), massive datasets are housed in a central, public-facing, cloud-based data store.

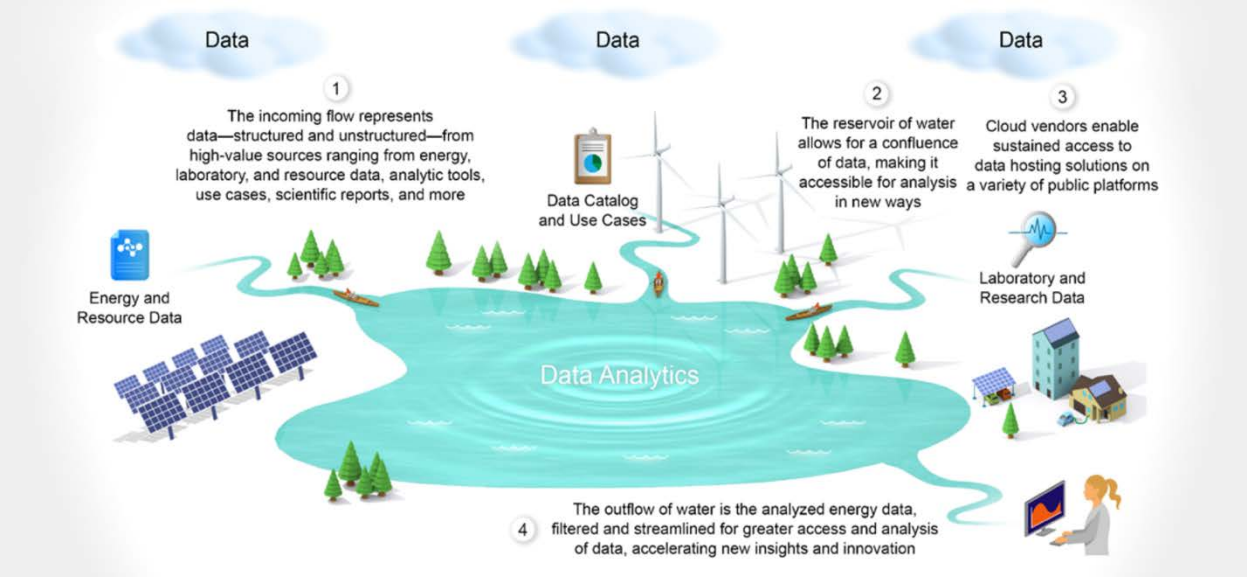

Figure 2. OEDI Data Lake illustration showing data flow and uses. Illustration by Besiki Kazaishvii, NREL.

Instead of downloading these data and performing research locally or at an institutional HPC solution to get answers to research questions, researchers send their research questions to the data in a data lake and need only download the answers. This can be done in several different ways, including encapsulating the research questions in modular code and pointing it at the data lake or by spinning up a server, or a cluster of servers, in an adjacent cloud environment that is

directly connected to the data. Throughout the process, the original data remain in the data lake providing consistent, equal access to the same dataset for all collaborators and eliminating the need for large data transfers between partners. By removing the expensive transfer component from the architecture, users of data lakes can accelerate their research timelines and save money.

The data lake paradigm also ensures consistent, universal access to the data. The reliance on a centralized data store ensures that potential collaborators have equal access to the same dataset without the need to transfer the data between partners, which has the substantial risk of data corruption during each transfer, or to the need to make the data operable within each institution HPC environment. Furthermore, cloud-based data lakes are accessible to anyone with access to the cloud, removing the requirement for a potential collaborator to have an HPC and big data storage solution.

4.1 GDR integration with OEDI

The GDR has been fully integrated with OEDI, which is currently home to over 128 terabytes (128 TB) of geothermal data, including trenched, or horizontal, DAS (DASH) and borehole, or vertical, DAS (DASV) SEG-Y and Hierarchical Data Format version 5 (HDF5) files containing data from the PoroTomo project. Data housed in an OEDI data lake appears in the GDR catalog as a regular resource (Figure 3) and is accessible via direct link to the data in the data lake on the applicable cloud provider as well as through an open data viewer, which allows users to browse the structure of the data within the data lake.

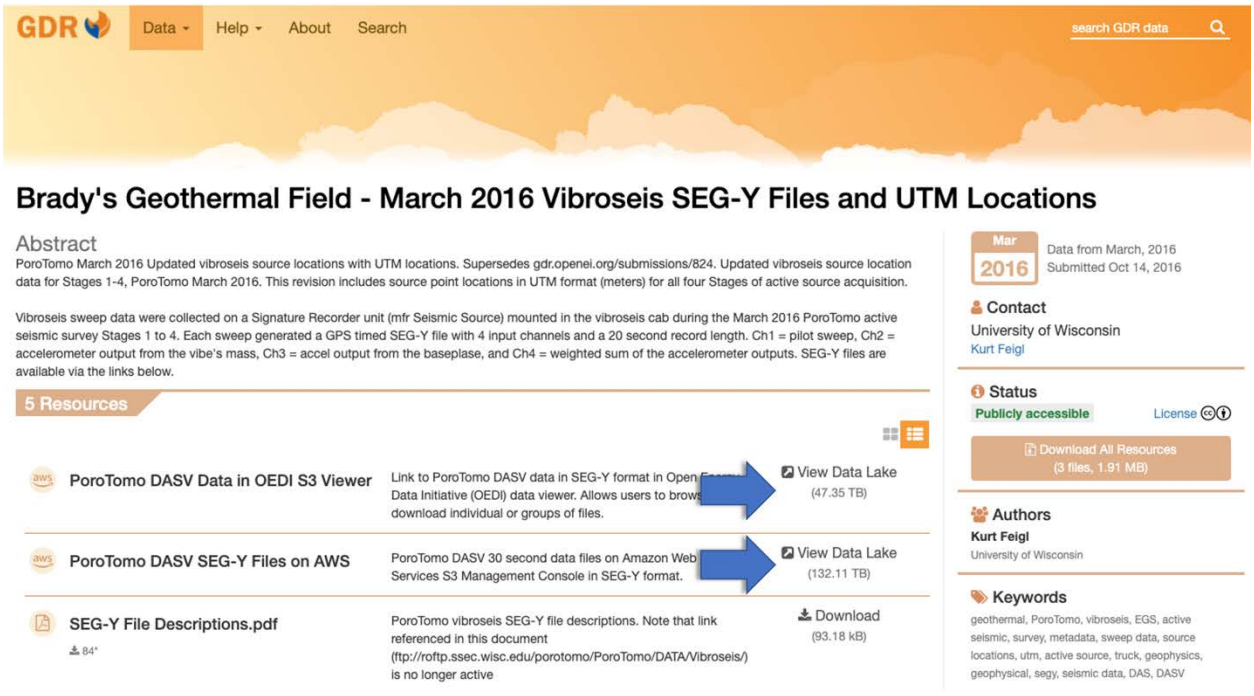

Figure 3. Screenshot of the PoroTomo Horizontal and Vertical DAS Data (Feigl 2017) on the GDR showing access to PoroTomo data in in SEG-Y and HDF5 formats via the "View Data Lake" links (blue arrows). This GDR submission is available at: http://gdr.openei.org/submissions/980.

By making these data available through the OEDI data lake, they are now accessible to anyone with access to the cloud. The data may now be processed by any organization, regardless of computational capabilities, removing previous barriers to use and making the data universally accessible.

4.1.1 Adding Big Data to the GDR

Data providers wishing to add a large dataset to the GDR should use the normal GDR data submission process to populate relevant metadata and either upload or link to any supporting files (e.g., data dictionaries, relevant publications, code, and contextual data), then contact the GDR team to discuss options for importing big data into the integrated OEDI data lake. The GDR team can be reached directly by emailing GDRHelp@ee.doe.gov.

5 Data Pipelines and Modern Data Standards

To improve the accessibility and usability of geothermal data and to provide support for larger datasets, the GDR has developed a new framework for the creation and application of data standards. Originating from modern programming trends, these new data standards are lightweight and have been built upon international data standards to support a variety of data types, including non-tabular, multi-dimensional data, time series data, images, and other formats incompatible with formats used previously to define data standards. Whenever possible, the standards are developed using Javascript Standard Object Notation with support for Linked Data (JSON-LD). This lightweight, machine-readable format supports the definition of multi-dimensional objects with nested properties and is supported by all modern programming languages. However, to avoid imposing limitations on data, the standards for each data type have been developed using the framework that best fits that individual data type and produces the most interoperable, universally accessible dataset. For some datasets, this may be simple Coma-Separated Values (CSV) format, while others may be best organized as a hierarchical format such as HDF5.

These new data standards are available on the GDR under the “Data” dropdown in the main navigation menu (Figure 4).

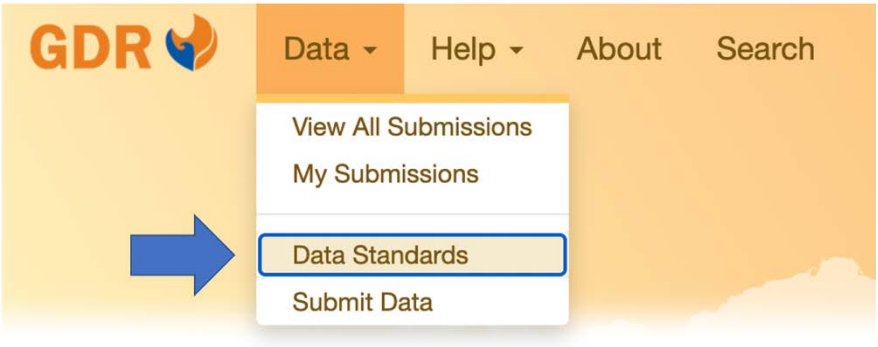

Figure 4. GDR main navigation showing "Data Standards" available under "Data" dropdown.

The identification of the optimal format and translation of data into that format can be an expensive process and can require the assistance of data scientists or programmers to implement. Rather than rely upon data submitters to perform these operations, the GDR has developed a series of data pipelines to automatically recognize and standardize popular data formats, e.g., drilling data.

5.1 Development of a Geothermal Drilling Data Standard

Recognizing the need for standardized drilling data, NREL has developed a drilling data standard, in partnership with DOE Geothermal Technologies Office (GTO). This standard has been documented and made available on the GDR under Data Standards (Figure 4) and was developed using the lowest denominator of common fields from common drilling data formats, specifically those deemed high value. While the standard is freely available for anyone to use, drilling data is often generated in formats proprietary or specific to the drilling rig or monitoring platform used by the drill operators. To make data from GTO-funded drilling operations as interoperable and accessible as possible, the data must be converted into the standard format. Instead of relying on operators to convert their drilling data or requiring individual research projects to undergo expensive data translation efforts, NREL has developed a data pipeline for the automated detection and conversion of drilling data and integrated it into the GDR at the point of submission.

Drilling data submitted to the GDR through the regular file upload and data submission process in a recognizable proprietary or nonproprietary format are automatically detected as drilling data and converted into the standard format (Figure 5).

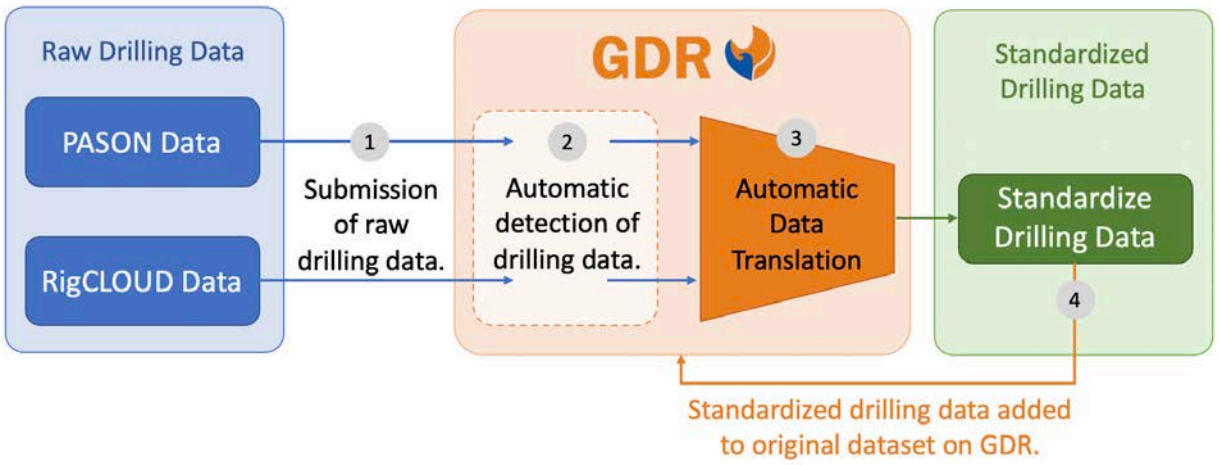

Figure 5. Data flow through a data pipeline for the automated detection and standardization of proprietary drilling data.

5.2 Standardizing Data Through Automation

Both the original, raw data and the converted standard data are made available to maximize the utility and accessibility of the submitted data. The GDR's data pipeline for automatic conversion of drilling data currently supports both Pason and RigCLOUD data formats. Additional guidance is made available for users of these platforms to configure their data acquisition platforms at the time of drilling to collect the proper fields and resolutions necessary to conform with the standard. This guidance can be found in the documentation of the drilling data standard, available on the GDR.

5.3 Broader Applicability

The implementation of data pipelines for data standardization has benefits beyond drilling data. Applicable to all forms of data, pipelines can be used to relieve data submitters of the burden of completing geothermal content models, allowing them to submit data in industry standard formats while still providing the greater research community with access to these data in standardized scientific formats. The GDR's logic-based approach can detect and process a variety of common data formats into their standard counterparts, and is capable of processing all forms of data, not just tabular data. Future data pipelines could include core photo metadata extraction and automatic association with drilling logs of shift reports, or the automated translation of Process Information (PI) from PI Systems into standardized operational data.

6 Conclusion

NREL's work to redefine data standards to support big data and non-traditional data types coupled with the development of data pipelines to automate the detection and conversion of popular data formats into these new standards enables cross-project analysis and generates high-value datasets capable of being used as the foundation for future data analysis activities or machine learning experiments while relieving individual research projects of the burden of data translation and conversion.

Additionally, the GDR has improved the accessibility and usability of geothermal data by integrating with OEDI and providing support for geothermal data lakes, making big geothermal data available to a broader audience. Access to these data is no longer limited to national labs, larger universities, and organizations that possess their own big data storage and compute infrastructure, enabling universal access to data for organizations of all sizes and opening the door to potential collaboration with smaller universities, high schools, startup companies, and other innovators.

Together these improvements enable equal access to energy data to support collaborative analysis and innovation across the greater geothermal industry.

Acknowledgement

This work was authored 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 with funding provided by U.S. Department of Energy Office of Energy Efficiency and Renewable Energy (EERE) Geothermal Technologies Office (GTO). The views

REFERENCES

Feigl, K. "PoroTomo Natural Laboratory Horizontal and Vertical Distributed Acoustic Sensing Data." [data set]. University of Wisconsin, Madison, WI (2017). Web. https://dx.doi.org/10.15121/1778858.

GDR. "DOE Geothermal Data Repository." OpenEI: Open Energy Information. National Renewable Energy Laboratory (NREL), 2 June 2021. Web. https://gdr.openei.org/.

Microsoft. "Excel specifications and limits." Microsoft Support. Microsoft, 1 June 2021. Web. https://support.microsoft.com/en-us/office/excel-specifications-and-limits-1672b34d-7043-467e-8e27-269d656771c3.

Weers, J., and Anderson, A. "The DOE Geothermal Data Repository and the Future of Geothermal Data." *Proceedings, 41st Workshop on Geothermal Reservoir Engineering*, Stanford University, Stanford, CA (2016).

Weers, J., and Anderson, A. "DOE Geothermal Data Repository: Getting More Mileage Out of Your Data." *Proceedings, 40th Workshop on Geothermal Reservoir Engineering*, Stanford University, Stanford, CA (2015).