The Geothermal Data Repository: Ten Years of Supporting the Geothermal Industry with Open Access to Geothermal Data

Jon Weers [(a)], Arlene Anderson [(b)], 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, repository, management, standards, dissemination, open, access, lake, GDR, OpenEI, DOE, OEDI, collaboration, big, storage, transfer, discoverability, usability, accessibility, pipeline, equality, provenance, innovation, cloud, submission, future, NREL

ABSTRACT

The Department of Energy's (DOE) Geothermal Data Repository (GDR) is celebrating its tenth anniversary! Over the last decade it has grown from the simple idea of storing public data in a centralized location to a valuable tool at the center of the US geothermal scientific community and an integral part of the DOE Geothermal Technologies Office (DOE GTO) project management strategy. Researchers funded by the DOE GTO have contributed over 1,300 data submissions to the GDR. These data have been used to further advancements in geothermal science, economic analysis, exploration, research, development, and operational efficiency. The adoption of open data methodologies and a data management strategy that prioritizes universal open access and standardized, interoperable data have further increased the value of GDR data, making them available across a distributed network of data sharing partners and improving their utility to other industries and related fields, including material science and space exploration. Incorporating feedback from users has been critical to the GDR's success, allowing it to grow over the years to meet the evolving needs of the geothermal community. This paper will explore some of many changes that occurred throughout the GDRs tenure and the lessons learned along the way, as well as highlight some of the new features and recent improvements that been

implemented to support innovation, reduce duplication of effort, and advance the geothermal industry as a whole.

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). The GDR was developed in 2012 by the National Renewable Energy Laboratory (NREL) in accordance with DOE's 2011 Strategic Plan, which stated that "DOE's success should be measured not when a project is completed or an experiment concluded, but when scientific and technical information is disseminated." The GDR was designed primarily to disseminate information and to protect DOE's investment in analysis, research and development activities by preserving data from those activities and communicating the resulting information to the largest audience possible. The metadata describing each dataset within 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.

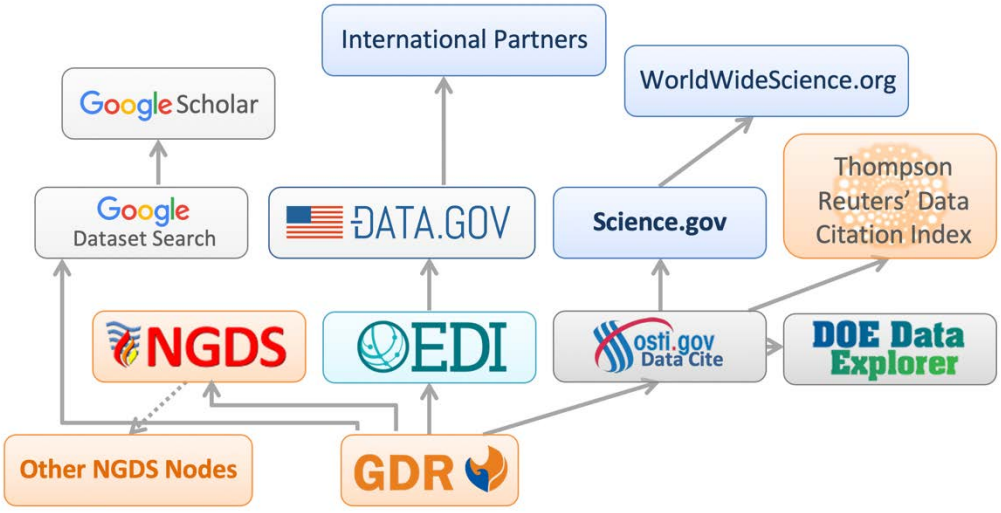

Figure 1. Diagram of GDR metadata propagation through a network of data sharing partners.

GDR data is discoverable and downloadable from each of the sites depicted above, without any duplication of effort or redundant storage. Requests for downloads from the data sharing network are routed behind the scenes to the GDR, enabling users to download GDR data without leaving the network partner's platform. By making data available throughout this network, the GDR has been able to increase the discoverability of DOE data by a factor of 12, when comparing direct downloads from the GDR itself to those originating from network partners (Weers et. al, 2021).

To date, the GDR has received 1,376 submissions and is now home to 5,084 resources and more than 135 TB of data from 74 different organizations (GDR 2022). Over the last 10 years, GDR data have been downloaded more than 3 million times by universities, national laboratories, private organizations, industry professionals, and government agencies.

The primary objectives of the GDR are to 1), protect DOE's investment in research and development by ensuring persistent, universal access to the results of GTO-funded activities; 2) overcome obstacles to collaborative data sharing; 3) collect and preserve as much data as possible by reducing the data management and submission burden for recipients of DOE funding, and 4) support innovation through the timely dissemination of DOE data.

2. Supporting Innovation through Agility and Adaptability

The GDR was specifically designed from day one to support innovation through the collection, storage, and dissemination of data resulting from research and development activities. This regularly required storing new forms of data, organizing data in new ways, and being able to accommodate new data formats and structures, and data representing entirely new technologies. The federation of metadata to the many sites in the network of data sharing partners (Figure 1) required the GDR to adopt numerous metadata standards. NREL developed a series of metadata translators that encompass the GDR at strategic endpoints to mimic the behavior of our target data sharing partners, allowing the GDR to transmit metadata in multiple standards simultaneously. To each of the sites in the data sharing network, the GDR appears to speak the required metadata language. The GDR is effectively using all the most popular metadata schemas simultaneously to ensure the broadest possible compatibility and to future proof data dissemination against metadata trends.

Additionally, the very framework of the GDR itself has been designed to be modular and adaptable. NREL utilized agile development methodologies and user-centric design to build a data repository capable of evolving to meet the ever-changing needs of the geothermal industry. The nature of supporting innovation requires the ability to quickly and easily expand to include or reorganize around new concepts and technologies.

2.1 Creating an Adaptable Framework

Support for research and development, especially in emerging fields, requires adaptability. An effective data management system must be able to receive and classify any type of data in its purview. Rather than require submitted data to adhere to a previously defined structure or metadata taxonomy, the GDR has employed a folksonomy (i.e. a user-generated system of classifying and organizing content). This allows the creators of data to introduce the classifications they deem most important, supporting the submission of raw data, new data formats, and the categorization of new technologies and concepts in real time.

Since the beginning, the GDR has been capable of accepting any type of data. There are no restrictions on the types of files that can be uploaded. Over time, the GDR's secure, cloud-based architecture was expanded to leverage advantageous native cloud technologies, such as infinitely scalable drives and data lakes, to remove limits on size and number of uploaded files. Today, the

storage capacity of the GDR automatically scales with each submitted file, ever expanding to perfectly meet the needs of the geothermal community.

2.2 Enforcing standards through curation

Data standards and metadata consistency are enforced through the curation process, in which the metadata for each data submission are reviewed by GDR curators for relevance, to ensure the metadata is descriptive of the data submitted; appropriateness, to make sure they are suitable for eventual public release; and completeness, to ensure no components are missing and sufficient metadata to promote discovery and proper use of the data. The original submitter's classifications are almost always preserved and are augmented by the addition of supplemental keywords and more verbose descriptions by the GDR curators. This approach has the added advantage of providing consistency throughout the data catalog while also accommodating innovative data structures, topics, and formats.

2.2 Responsive to Users' Needs

For the entirety of the GDR project, NREL has utilized user-centric design and development methodologies to be responsive to industry needs in a timely manner. Regular attendance at conferences, quarterly trainings, help desks at major industry events, and other user outreach programs have allowed the GDR team to collect valuable feedback from users including suggested improvements, changes in data submission and publication workflows, and strategic additions to the GDR's network of data sharing partners. These improvements are prioritized and included in regular, iterative development cycles and are often released as new features or improvements to the GDR.

Additionally the GDR teams provides technical assistance to data submitters, assisting them with everything from data organization and data management planning at the beginning of a project to data standardization and large file transfers at the time of submission.

2.3 Encouraging Data Citation and Provenance

For many years now, the GDR has been at the forefront of the open data movement for DOE. In addition to preserving and disseminating DOE data, the GDR team actively encourages the proper use and citation of datasets to support scientific posterity and provenance. License to use information is prominently displayed on every GDR dataset. All eligible datasets receive a Digital Object Identifier (DOI) from DOE's Office of Scientific and Technical Information (OSTI), and the metadata for each dataset is pre-formatted into a popular citation format on every dataset landing page for easy inclusion in publications. The proper citation of data will enable provenance from source to finished product, especially in the case of derivative data products such as modeling or analysis efforts, or data amalgamations. Recently, the GDR citation generator was expanded to include all popular citation formats (Figure 2), including MLA, APA, Chicago, and RIS; allowing for each to be copied to the user's clipboard with a single click.

Figure 2 The citation selector from a GDR dataset showing the MLA formatted citation.

3. Lessons learned

Not everything that exists in the GDR today was conceived of ten years ago. Some of the more impressive features came about through years of iteration, refactoring in response to user feedback, or incorporating lessons learned and data management success stories from other industries. Several valuable lessons learned have paved the way for automation, smarter submission forms, improved training, and a new paradigm for data standardization.

3.1 Automation and Simplification

One of the more significant improvements of the last decade has been the transition toward an automated, simplified data submission process. To meet the needs of users at the time it was created, the GDR originally supported 3 distinct and separate data submission pathways, including a web form and the ability to upload an Excel metadata template (Weers and Anderson, 2013). This proved to be confusing to users and costly to maintain as improvements had to be made to all three data ingress methods for consistency. The NREL team transitioned the GDR to a singular submission pathway and allocated resources saved on maintenance towards the automation and streamlining of the submission process. The development of a smart form made the process even easier for data submitters by automatically populating key metadata fields based on intelligent assumptions, prior submissions, detected file types, and other input factors. For example, users submitting open-source code as part of a data submission need only link to the GitHub repo address. The GDR will automatically reach out to GitHub and import any associated metadata, populating the required metadata for the code submission automatically and in most cases, without any further interaction from the submitter. The development and refinement of the smart form makes submitting data easier, results in better quality metadata, and reduces the burden placed on data submitters, allow more DOE resources to be spent on research and development activities.

3.2 Improved Training

NREL has always strived to create an intuitive interface for the GDR and to minimize the amount of training necessary. However, the GDR team quickly realized that proper data management extends beyond the GDR application and that quality data begins at project inception. Early training sessions were modified to include guidance on the organization and management of data throughout the project lifecycle to ensure that sufficient metadata were collected along the way to support the eventual submission of project data to the GDR. These

long, comprehensive training sessions were often done as a quarterly webinar or conducted on-demand for new DOE funds awardees. A recording of the training session was made available on the OpenEI YouTube channel and, in response to user feedback, a written version was created for those that prefer to read best practices documents. Eventually, a series of short, targeted micro-video tutorials (no more than 1-2 minutes) were created, allowing the GDR team to go into more depth on each individual data submission field without overwhelming general audiences. These videos have been organized into a GDR tutorials playlist on YouTube.

Leading by example, the GDR team recently created GDR submission #1 to house and showcase data management best practices for GDR submitters and curators (Weers et al, 2021). The submission contains all the resources listed above and serves and an example submission to others.

3.3 Evolution of Data Standards

One of the greatest lessons learned in the last ten years has led the GDR team to revisit the practices around data standardization. Originally, data submitters were encouraged to submit standardized structure data, in the form of content models created by topic for use across the National Geothermal Data System (NGDS), of which the GDR is a node. While these standards are still supported by the GDR, their adoption by data submitters has been sparse and users have complained that the required formats (XML or Excel spreadsheet) are limited and don't allow for the standardization of non-tabular data formats such a time-series data, geospatial data, or multi-media formats such as video files. Furthermore, the requirement for data submitters to translate vast amounts of data from industry standard formats such as SEG-Y into NGDS data standards proved to be costly and time consuming.

The GDR team's new approach to data standardization takes advantage of modern data formats such as JSON-LD and builds upon the successful standardization efforts and data management best practices used by the banking and medical industries. Rather than require awardees to translate data into standard formats, the GDR will now utilize data pipelines to detect select data in its original form and automatically standardize it. For example, drilling data submitted to the GDR are automatically detected as drilling data and converted into a standard format (Figure 3).

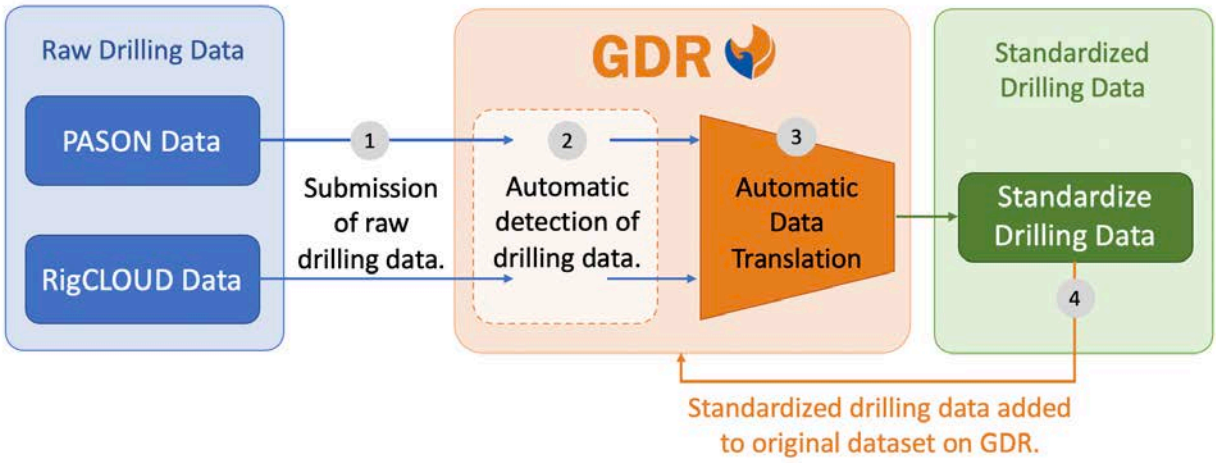

Figure 3. Data flow through the GDR data pipeline for the automated standardization of drilling data.

This approach promotes data sharing by allowing submitters to submit data in whatever formats they are generated, even if the formats are raw or proprietary, transferring the burden of standardization from individual project teams to the GDR's team of data experts, resulting in reduced duplication of effort and greater efficiencies across the DOE portfolio and allowing more resources to be spent on critical project objectives.

4. Support for Big Data Through Data Lakes and Integration with the Open Energy Data Initiative (OEDI)

The Open Energy Data Initiative (OEDI) provides universal access to high-value datasets through cloud-based data lakes to accelerate analysis and advance innovation. NREL has partnered with major cloud providers, including Amazon Web Services (AWS), Google Cloud, and Microsoft Azure to make more than 1 petabyte (1 PB) of high-profile data universally accessible to the public. The GDR was integrated with OEDI to create a Geothermal Data Lake (Weers et al 2021) to overcome the challenges of working with big data and make big geothermal data available to a broader audience.

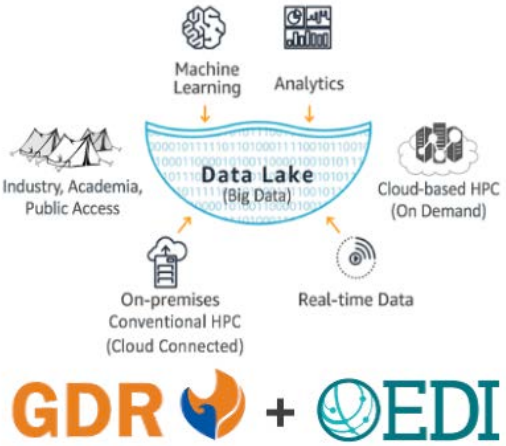

Figure 4 GDR integration with the Open Energy Data Initiative (OEDI) to produce a geothermal data lake

In the past, large datasets had to be downloaded or otherwise transferred to an organization with a High-Performance Compute (HPC) capability where researchers would analyze their findings and run analyses to get answers to their research questions. In the data lake model, researchers send their research questions to the data and only download answers. There are numerous ways to connect to a data lake, represented by the different arrows in Figure 4, including encapsulation of the research question in a container, the use of modular code, the development of a cloud based HPC by spinning up new servers, or clusters of servers, in an adjacent cloud, or through connected applications and support tools such as Jupyter notebooks. Data in the data lake remain intact, available to all research partners, enabling real-time collaboration on big data without the need for large data transfers between partners. Users of data lakes can accelerate their research timelines and save money by removing the need for costly data transfers and synchronization among partners. The use of a centralized data store also ensures consistent access to the data for all collaborators, reducing the risk of data corruption and the duplication of effort needed to make the data operable at each institution.

Figure 5 shows a portion of the PoroTomo GDR submission, which includes links to the PoroTomo distributed acoustic sensing (DAS) dataset hosted by the OEDI data lake. This submission provides access to the data in a more traditional industry standard format (SEG-Y) along with a cloud-optimized hierarchical data format (HDF5). It also includes a tutorial Jupyter Notebook which allows the user to walk through an example of accessing the data and applying some basic processing techniques within a Jupyter Notebook, all without downloading any data. The PoroTomo DAS data has quickly become one of the GDR's most popular datasets due to it being of high value for research in DAS applied to geothermal resource characterization, acting as one of the first cloud-accessible DAS datasets, and providing an example of how to access data stored in the cloud.

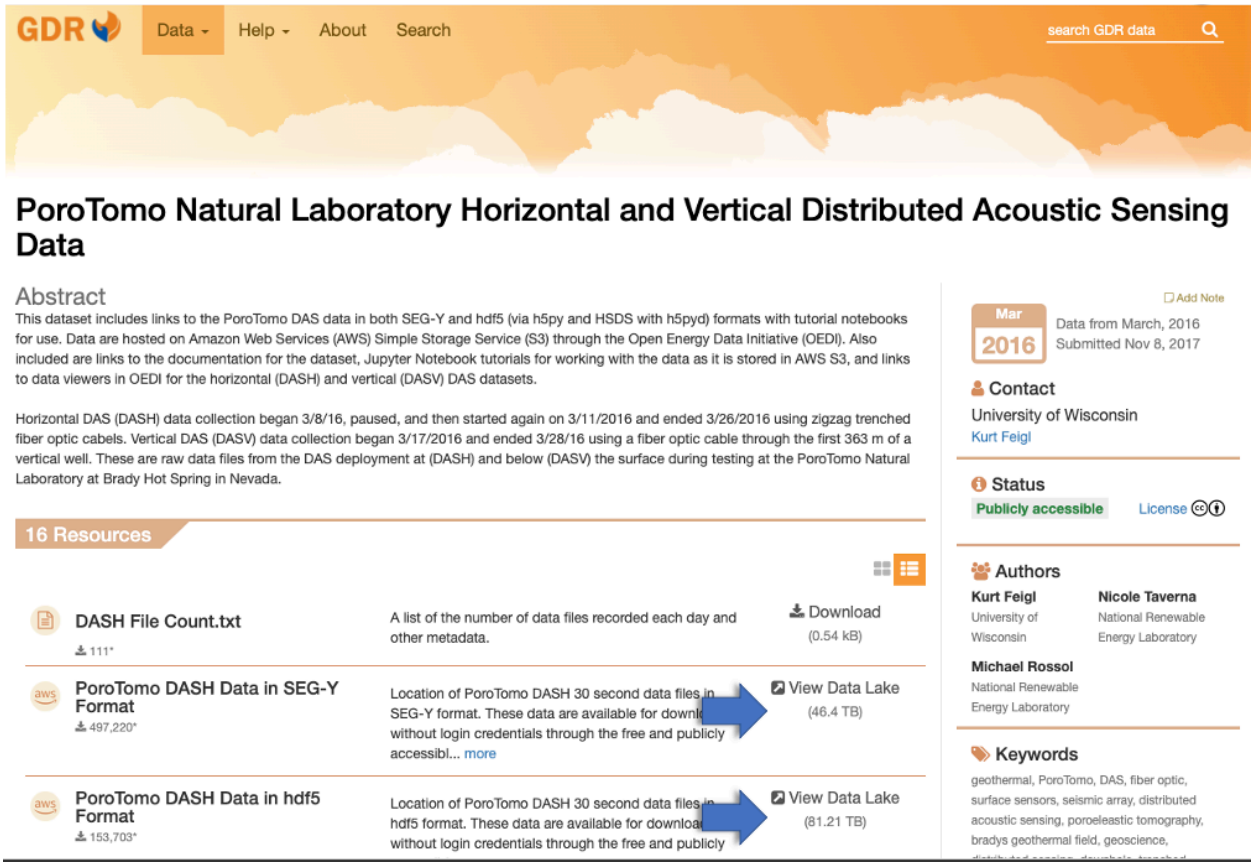

Figure 5. Screenshot of the PoroTomo Horizontal and Vertical DAS Data (Feigl 2017) on the GDR showing access to PoroTomo data in in SEG-Y and HDF5 formats via the "View Data Lake" links (blue arrows). This GDR submission is available at: http://gdr.openei.org/submissions/980.

Access to data in the GDR data lake are no longer limited to national labs, larger universities, and other organizations that possess their own big data storage and HPC 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, underprivileged communities and other innovators.

5. Overcoming the Challenges of Data Sharing

There are many challenges to creating a culture of proper data management and collaborative data sharing (Figure 6), including technological (e.g. transfer speeds and storage limits), organization (e.g. proprietary data, institutional policy), financial (e.g. cost of infrastructure needed to support big data), and cultural (e.g. believe that my data is my asset). The GDR team has experienced these challenges first-hand, and has employed a several approaches to combat them, including a state-of-the-art cloud architecture, infinitely scalable drives, integration with the Open Energy Data Initiative (OEDI), the formation of a geothermal data lake, and annual training for awardees on data management best practices and data sharing success stories.

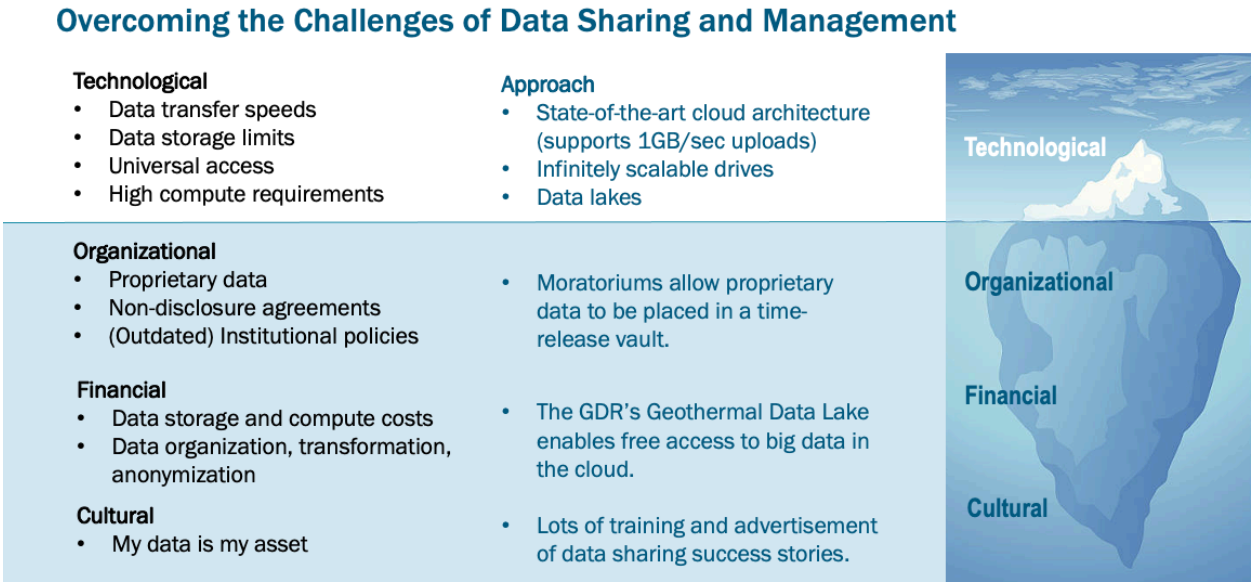

Figure 6 Challenges of data sharing and the approaches taken by the GDR team to overcome them

The positive impact of the GDR team's efforts to overcome the challenges of sharing data can be quantified through the number of downloads (over 3 million), datasets (1,019) , and volume of submitted data (135.28 TB), but just as notable are some the qualitative metrics and success stories captured over the years.

5.1 Success Stories

The evolution of the questions emailed to the GDR team are one such metric, and over the years paint a positive trend and a notable impact in the effort to overcome organizational and cultural challenges to data sharing and management. Many early questions received around the time the GDR was created, in 2012 and 2013 were focused on trying to understand data submission requirements and addressing perceived stigmas with sharing data. Questions such as, *"Why do I have to submit my data?"* and *"What data does DOE want?"* were prevalent. The concerns were addressed through hosting workshops and training sessions, by placing help desks at major industry functions, by handing out awards at DOE events for best data submission and most downloads, through technical assistance and one-on-one consultations, and through the

communication of data management best practices and data sharing success stories. Over time, as the value of sharing data was demonstrated more successfully, the questions received by the GDR team switched from inquiries of obligation to inquiries of potential. It became commonplace for the GDR team to receive questions like, *"Can you handle 10 TB of data?"* and *"What is the best format for time-series data?"*. These questions not only demonstrated an enthusiasm for the GDR platform, but they also helped shape the future of GDR development and the allocation of resources towards support for big data and new data standards. In recent years, support questions have become even more flattering, with users asking if they can amend previous submissions with updated information, asking us to help coordinate the assignment of DOIs with the release of a new publication, or even asking if their personal, non-DOE-funded research projects can be added to the GDR so that they might benefit from the GDR's powerful dissemination engine.

5.1.1 Demonstrating the Value of Raw Data

One of the most significant success stories of the last ten years is the quantifiable impact of the potential for raw data to be reused in new and exciting ways. Throughout the last decade, GDR training materials have touted the intrinsic value of raw data as a fuel for innovation. Summarizing project data is, by its very nature, biased towards the scope and objectives of the project. While raw data remains free to be used and reused in any number of different ways. The best example of this from the last 10 year shows up as the GDR's second most downloaded dataset of all time: the Evaluation Data of a High Temperature COTS (Commercial Off-the-Shelf) Flash Memory Module (TI SM28VLT32) for Use in Geothermal Electronics Packages (Cashion 2014). Downloaded over 95,000 times in total, the dataset frequently takes the #1 or #2 spot in the top 10 downloads per month (Figure 7), prompting the GDR team to dig a little deeper into the demand for this data.

Top 3 Most Downloaded Datasets for March 2022

Rank	Downloads	Submission Name	Size
1	29,980	PoroTomo Natural Laboratory Horizontal and Vertical Distributed Acoustic Sensing Data	132.11 TB
2	19,947	Evaluation Data of a High Temperature COTS (Commercial Off-the-Shelf) Flash Memory Module (TI SM28VLT32) for Use in Geothermal Electronics Packages	284.53 MB
3	18,781	Brady's Geothermal Field Nodal Seismometer Data	2.2 TB

Figure 7 Table showing top 3 most downloaded GDR datasets for March 2022.

Looking for a trend to explain the consistently high volume of downloads, the GDR analysts were expecting to see evidence of a university using the dataset for a class project, a poorly

written script automating download inside a loop, or some other explanation for the astronomical access numbers. However, what they found was a long list of legitimate downloads from tens of thousands of separate individuals from as many different organizations, ranging from universities to private corporate research firms, many of them operating in industries unrelated to geothermal, including researchers from aerospace, the automotive industry, and cell phone manufacturers. The conclusion was that the raw data from of the evaluation of memory modules at high temperatures was valuable to anyone working with electronics in potentially hostile environments.

6. Conclusion

Over the last 10 years, the DOE and NREL have worked together to ensure that the GDR remains at the forefront of the DOE open data movement and that it continues to meet the data management and information dissemination needs of the geothermal community.

Looking forward, the GDR team aims to expand data pipeline efforts to automate the standardization of additional data types, develop value-add data products and aggregations, support innovation, reduce duplication of effort, and advance the geothermal industry.

The GDR protects DOE's investment in research and development by employing an adaptable approach to data management that enables the GDR to provide open, transparent access to public data; security for restricted access data; data provenance and preservation for the geothermal scientific community; and to be ready to support whatever advancements and innovation the future of geothermal may hold.

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 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

Cashion, A. "Evaluation Data of a High Temperature COTS (Commercial Off-the-Shelf) Flash Memory Module (TI SM28VLT32) for Use in Geothermal Electronics Packages." [data set]. Sandia National Laboratories (2014). Web. https://dx.doi.org/10.15121/1154909.

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), 25 April 2022. Web. https://gdr.openei.org/.

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).

Weers, J., and Anderson, A. “Fueling Innovation and Adoption by Sharing Data on the DOE Geothermal Data Repository Node on the National Geothermal Data System.” *Proceedings, 38th Workshop on Geothermal Reservoir Engineering*, Stanford University, Stanford, CA (2013).

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., Taverna, N., Huggins, J., Scavo, RJ. “GDR Data Management and Best Practices for Submitters and Curators” [data set]. National Renewable Energy Laboratory, Golden, CO (2021). Web. https://gdr.openei.org/submissions/1.