Connecting People to Data: Enabling Data Connected Communities through Enhancements to the Geothermal Data Repository

Jon Weers [(1)], Arlene Anderson [(2)], and Nicole Taverna [(1)]

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

15013 Denver West Parkway

Golden, CO 80401-3305

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

1000 Independence Ave. SW

Washington D.C. 20004, USA

Keywords

Geothermal, data, repository, management, improvements, community, people, industry, engagement, features, enhancements, open, access, GDR, OpenEI, DOE, collaboration, discoverability, usability, accessibility, innovation, submission, metadata, machine learning, ML, large language model, LLM

ABSTRACT

The Department of Energy's (DOE) Geothermal Data Repository (GDR) has implemented a series of new features designed to connect people to data. These features, which are based on feedback from the GDR user community and surveys of the greater geothermal research community, are designed to improve data quality and empower members of all communities to better engage with geothermal data resources by providing universal access to data and by improving the connections between data providers, subject matter experts, and the communities of people using GDR data. This paper will explore some of the recent enhancements made to the GDR to improve data discoverability, reduce submission time, and result in better quality data submissions. These improvements include the ability for users to save a list of their favorite datasets, search for insight into geothermal datasets or data availability, or sign up to receive notifications of future updates to specific datasets. These improvements aim to enhance the overall user experience of the GDR while further connecting communities to the data they need to inform decisions, advance geothermal research, and develop innovative solutions to local energy problems.

1. Introduction

Developed by the U.S. Department of Energy (DOE) to receive, manage, and make available all relevant data generated from projects funded by the DOE Geothermal Technologies Office (GTO), the Geothermal Data Repository (GDR) has been designed since its inception to be a resource for the geothermal community. Its primary objective is to protect DOE's investment in research, analysis and development through the proper management of data and information and dissemination to the public to fuel innovation, reduce duplication of effort, and promote scientific discovery in geothermal technologies. To be successful in reaching its objective, the GDR has to be used by the geothermal community. The GDR team has adopted an iterative, agile development methodology and developed numerous features and enhancements to the GDR over the years to ensure that it continues to be a useful tool in support of the geothermal community.

To date, the GDR has received 1,516 data submissions and is now home to 5,484 resources and more than 140 TB of data from 107 different organizations (GDR 2023). GDR data have been downloaded more than 3 million times by the greater geothermal community, including various academic institutions, national laboratories, private organizations, industry professionals, and government agencies.

2. Community Informed Design

The GDR is a tool built for the geothermal community with direct input from the geothermal community. Throughout its tenure, the GDR team has worked to overcome many obstacles to data sharing to build a strong data sharing culture within the geothermal community (Weers et al, 2022). To be successful, the GDR must be useful for the geothermal community. One of the easiest ways to do that is to involve the community in the design and development process. This requires both listening to feedback from the community as well as adopting a development methodology that allows for that feedback to be incorporated.

2.1 Listening to the Community

Listening to feedback and incorporating it into future development activities is a key part of the GDR team's strategy to ensure that the GDR continues to meet the needs of the geothermal community. Feedback is collected annually from a wide array of stakeholders, not just from traditional IT-related data communities and GDR users, but also from geothermal experts, industry professionals, academic institutions, and geographic communities. For example, GTO is funding several coalitions of stakeholders to design and deploy community geothermal systems (GTO 2023). These coalitions are planning to work together on developing case studies and data that will be included in GDR's Low-Temperature sections, including Geothermal Heat Pumps (GHPs), Geothermal District Heating and Cooling (GDHC), and Direct Use applications. The GDR team will work closely with these coalitions, providing trainings on data management and submission best practices tailored to their community and soliciting their feedback to ensure that the GDR is meeting their specific needs as well as the needs of the greater geothermal community.

2.2 Agile, Iterative Development

An effective data management system must be adaptable. This is especially true in the case of the GDR, which supports research and development activities, including projects working with emerging technologies and novel approaches (Weers et al 2022). These innovations often require modifications to data classifications, metadata standards, and other GDR infrastructure designed to make the data more discoverable. From the start, the GDR team adopted an agile software development methodology to allow the GDR to easily accommodate community feedback and changes to support new technologies at any point in the development timeline (Figure 1).

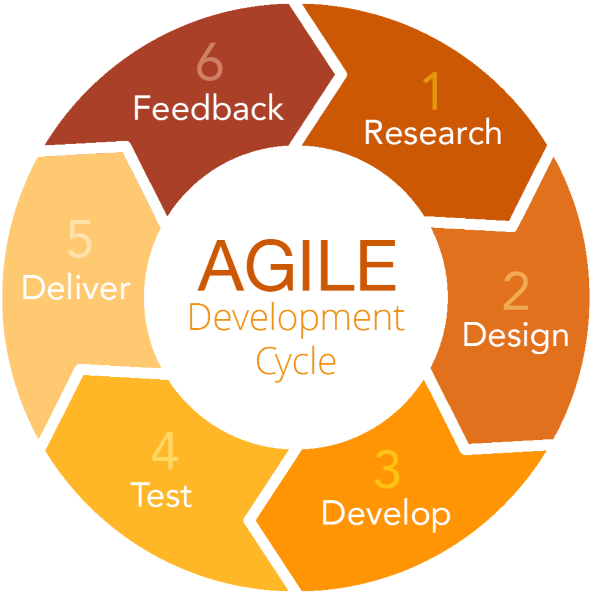


Figure 1 Agile Development Cycle showing the steps within each iteration

Agile software development is a process in which development is divided into a series of short iterations. Each iteration includes a design session that allows the development team an opportunity to review the tasks at hand and reconcile them with the current state of the tool, its community of users, and any feedback received from stakeholders. Larger tasks are subdivided into smaller tasks so that no single task takes longer than the iteration. Those tasks are then developed, tested, and deployed by the end of the iteration to get hands-on feedback from users and other stakeholders in time to inform the next iteration. This iterative process allows the development team to re-evaluate the trajectory of larger efforts at regular intervals and adapt them as needed to ensure that every development effort continues to support the community's needs, even as those needs evolve.

The GDR has been utilizing this agile development methodology for over ten years, allowing it to evolve over time along with and in support of the geothermal community (Weers et al 2022).

3. Connecting People to Data

Several new features have been added to the GDR, based on feedback from the GDR users and members of the larger geothermal research community. These features have been designed to better connect members of the community with relevant data, provide additional insight into GDR data resources, and improve connections between data providers, subject matter experts, and the communities of people using the GDR. Some of most requested features have centered around enabling user to interact with datasets and their creators to ask questions of the data, leave comments, favorite or rank certain datasets rank, and/or subscribe to update notifications for a particular dataset. Many of these features have recently been implemented in the GDR and their details appear below. Others, such as the ability to comment on a dataset, pose interesting challenges that could potentially be resolved through the adoption of emerging technologies, including the adoption of a custom-trained large language model (LLM).

3.1 Stars and Subscriptions

The GDR Team has recently received a lot of requests for the ability to rank datasets, favorite certain datasets, and to subscribe to update notifications for select datasets. These concepts are not new and there are many successful examples of them implemented across similar tools on the internet. After careful consideration, the GDR team narrowed down potential solutions to two models, which, for the purposes of this paper, are named after well-known sites that have adopted (and arguably even perfected) their use.

3.1.1 The Stack-Overflow Model

Stack Overflow (Stack Exchange 2023) is a well-known developer resource and knowledge base housing questions and answers that allows users to “upvote” or “downvote” answers to a given question, creating a system by which the best answers rise to the top. Stack Overflow also allows users to “bookmark” their favorite answers to find them easily in the future. This model was considered both as a means for managing inappropriate comments (discussed later) and a way of ranking datasets.

3.1.2 A Careful Calculus

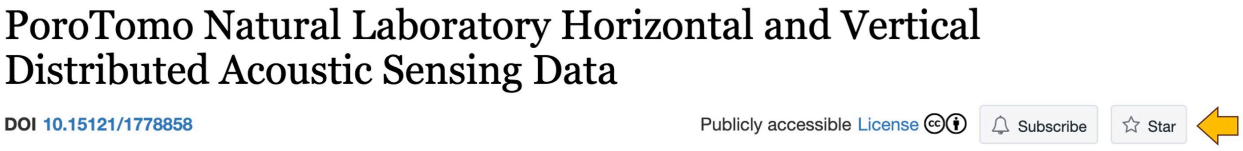
The GDR team ultimately decided against adopting the Stack Overflow model over concerns that ranking geothermal datasets was fundamentally different than ranking attempts to answer the same question. GDR data originates from GTO-funded research and development activities, which by their very nature, are heterogenous, making them difficult to compare to one another. Each funded activity typically aims to solve a unique problem, test a novel theory, or otherwise perform work likely to produce a unique dataset. The GDR team felt a quantitative ranking system would do many datasets injustice as any applied rankings could potentially lead to false comparisons. For example, the ranking of data from two different siting projects could be misconstrued to be an indicator of each site’s geothermal potential instead of the quality of the dataset itself.

3.1.3 The GitHub Model

GitHub (GitHub 2023) is a well-known code repository and catalog of open-source code for developers and software engineers that contains many different codes supporting almost as many different projects. GitHub catalog users can favorite a code repository by clicking a “star” icon, subscribe to update notifications by clicking a “bell” icon, or both. This simple approach seemed to adequately address the needs of users to track their favorites without inviting direct comparisons or insinuating quality. Additionally, the content of the GitHub catalog, which consists of unique code repositories serving a multitude of purposes, is more analogous to the content of the GDR. For these reasons, the GDR team chose to adopt this model.

3.1.4 The GDR Implementation

Modeled after GitHub, the GDR now allows users to both “star” and “subscribe” to a dataset by clicking either the “star” icon or “bell” icon directly underneath the dataset title (Figure 2).



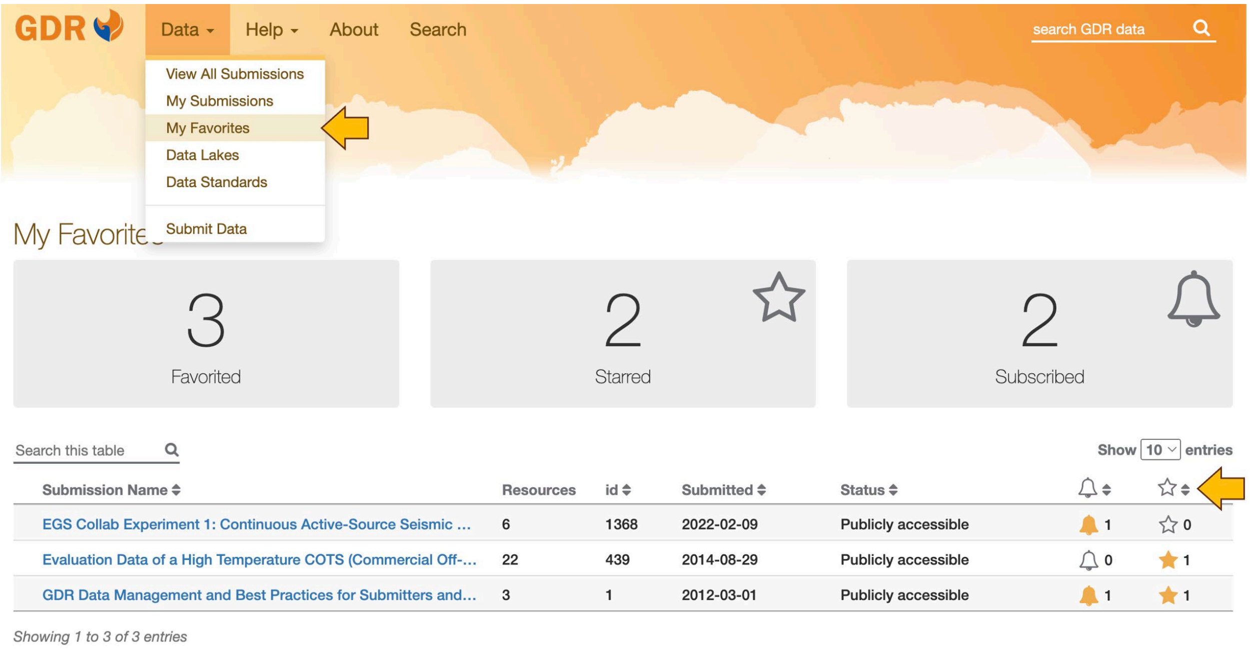
PoroTomo Natural Laboratory Horizontal and Vertical Distributed Acoustic Sensing Data

DOI 10.15121/1778858

Publicly accessible License

Figure 2 Screenshot of the new "Subscribe" and "Star" functionality on GDR Datasets.

Clicking these icons a second time will “unstar” or “unsubscribe” the user. Users can also manage their subscriptions and stars by accessing the “My Favorites” feature under “Data” (Figure 3, top left arrow), which will display all the submissions the user has starred or subscribed to with the option to sort the list by stars, subscriptions, name, id, submission date, status, or to search the list by keyword or phrase (Figure 3, bottom right arrow).



My Favorites

3 Favorited

2 Starred

2 Subscribed

Submission Name	Resources	id	Submitted	Status	🔔	☆
EGS Collab Experiment 1: Continuous Active-Source Seismic ...	6	1368	2022-02-09	Publicly accessible	1	☆ 0
Evaluation Data of a High Temperature COTS (Commercial Off-...	22	439	2014-08-29	Publicly accessible	0	★ 1
GDR Data Management and Best Practices for Submitters and...	3	1	2012-03-01	Publicly accessible	1	★ 1

Showing 1 to 3 of 3 entries

Figure 3 Screenshot of the GDR's "My Favorites" page with options to search and sort favorited datasets.

This simple approach is easy to understand and provides users with the desired functionality while also serving as a rudimentary ranking system. Datasets which have received many stars can be prioritized in search results and the number of stars a dataset receives can serve as a qualitative measure of the datasets impact without invoking a direct comparison to other datasets. Future development activities involve refining the GDR search interface to include an option to order search results by number of stars.

3.2 The Problem with Comments

Another frequently requested feature is the ability for users to comment on datasets. The interesting thing about this feature request is that at one point the GDR actually supported this functionality. GDR users used to be able to post comments on a dataset's page, but the feature was subsequently removed at the request of both users and DOE due to the proliferation of irrelevant and inappropriate comments. The GDR team quickly found that the majority of comments posted on dataset pages were either textbook spam (e.g. ads selling scams or illicit substances) or requests for information that were largely unrelated to the dataset itself, including requests for assistance installing software and requests for additional data unrelated to original project (e.g. *"This geothermal dataset is great! Do you have any transmission data?"*).

3.2.1 Lessons Learned

The original GDR comments functionality started with the best of intentions: to allow users to post questions about a dataset and for subject matter experts to post useful answers. However, supporting this functionality required a tremendous amount of maintenance. New comments, which came in by the dozens daily, had to be screened for appropriateness and relevance. The majority of them (over 95%) were discarded during the screening process and the few that remained added little value to the original datasets. It became quickly apparent that in order to make the comments valuable, the community would have to be enlisted to help police the comments and responses. This would have required the development of a comment moderation system that included the review and curation of submitted comments, managing the review queue, assigning moderators, and ranking the relevance of approved comments. This is a lesson learned all too well by the team at Stack Overflow, who has adopted all these things and more. Community moderators at Stack Overflow police new comments as they are added. These moderators are volunteers elected by the community to serve annual posts. An entire system has been constructed to incentivize and reward Stack Overflow's army of moderators, including badges, digital trophies, a reputation system and other forms of recognition (Stack Overflow 2023). It's a wonderful example of gamification to entice participation in a complicated system of community moderation, but also a good example of much supporting infrastructure can be necessary to have an effective comments section.

In the end, the GDR team elected to remove the ability for users to post comments on datasets. However, in light of recent, renewed interest from the community, that functionality is being considered once again.

3.2.2 Digging Deeper

In an effort to better understand the motivation behind recent requests for comments to be added to the GDR, the team reached out to several of the requestors for additional insight. The underlying requests appear to be centered around the idea of having comments serve as a frequently asked questions (FAQ) database of sorts with both questions and answers related to the dataset. This would serve two primary functions:

1) Reducing the number of repeat questions asked of a dataset's primary point of contact.
2) Providing data users with quick and easy access to contextual and supporting information for a dataset without having to wait for a dataset's primary point of contact to respond.

Fortunately, these two functions may be able to be served through the adoption of a new, emerging technology.

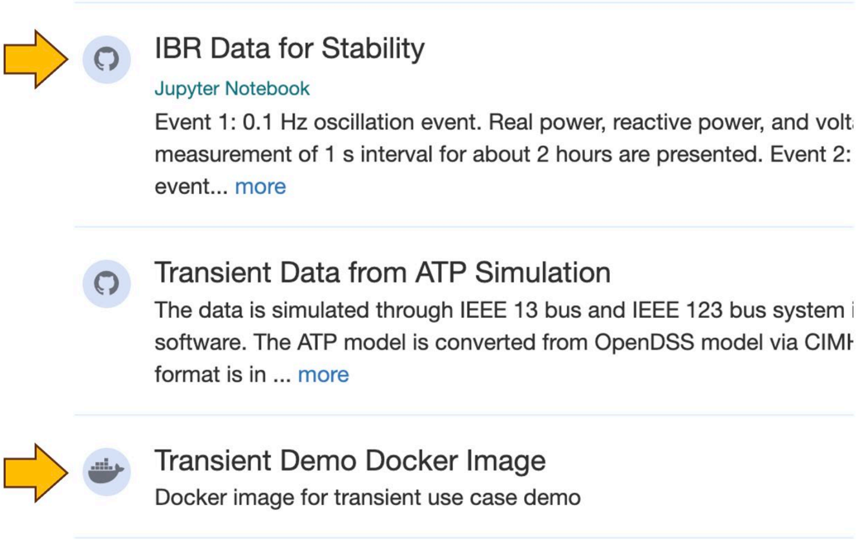
3.2.3 Potential for Better Insight Through Use of Emerging Technology

The GDR team is exploring an alternative to a traditional comments section using an adaptation of machine learning (ML) and large language models (LLMs) like ChatGPT. By successfully training an LLM on the corpus of knowledge contained within the GDR, including the metadata provided with each dataset and the content of associated supporting documents, including conference papers and journal articles, the GDR team could develop a chat-based tool or intelligence search interface capable of answering most of the questions that would be posted in a comments section. These answers could be limited to the knowledge contained within the GDR and related publications, reducing the likelihood of false associations and improving the quality of answers over traditional LLMs. While an LLM likely couldn't extrapolate complex answers (e.g. any answers involving calculations), it could easily provide answers covered in supporting materials. In addition, any answers provided could include citations to the source materials used to generate the answers, including referenced publications and GDR datasets.

Early prototypes have been very promising and have demonstrated a potential to provide both of the primary functions outlined above: 1) reducing the number of questions asked of dataset contacts and 2) providing users with timely answers to their data questions.

3.4 Power to the People

Additional improvements have been made to the GDR, including the ability to edit links after a dataset has been published, the automatic recognition of resources pointing to Jupyter notebooks and Docker containers (Figure 4), and the addition of direct access links for data lake resources (Figure 5).


IBR Data for Stability

Jupyter Notebook

Event 1: 0.1 Hz oscillation event. Real power, reactive power, and volt measurement of 1 s interval for about 2 hours are presented. Event 2: event... more

Transient Data from ATP Simulation

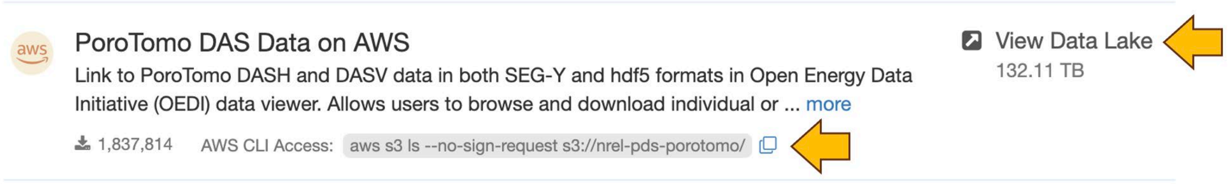
The data is simulated through IEEE 13 bus and IEEE 123 bus system i software. The ATP model is converted from OpenDSS model via CIMI format is in ... more

Transient Demo Docker Image

Docker image for transient use case demo

Figure 4 Screenshot of a dataset containing both a Jupyter Notebook link (top arrow) and a link to a Docker image (bottom arrow).

Links to Jupyter notebook or Docker container resource will be automatically detected on during data submission, and any supporting metadata already hosted on public repositories, like GitHub or DockerHub, will be automatically ported into the data submission by the GDR, saving the submitter from having to input the same information twice. The attributed metadata is, of course, editable, should the submitter wish to make any changes or corrections to the information automatically imported.



PoroTomo DAS Data on AWS

Link to PoroTomo DASH and DASV data in both SEG-Y and hdf5 formats in Open Energy Data Initiative (OEDI) data viewer. Allows users to browse and download individual or ... more

View Data Lake
132.11 TB

1,837,814 AWS CLI Access:

```
aws s3 ls --no-sign-request s3://nrel-pds-porotomo/
```

Figure 5 Screenshot of a data lake resource with links to browse the data through the data lake viewer (top arrow) and to copy the code for direct command line access to the data (bottom arrow)

Previously, data lake resources had two separate links: one to a catalog entry in the cloud provider's public datasets program, which contained information on connecting directly to the data, and one to the Open Energy Data Initiative (OEDI) a data lake viewer, which allowed users to browse data within the data lake and download sample data files. These two resources have been combined (Figure 5) and links to the data lake viewer now also contain a direct access link and a convenient quick copy button (Figure 5, Orange Arrow) that copies the command line code needed to access the data into the user's clipboard.

Together, these features empower both novice and expert users by providing quick and easy access to GDR data through both a web-based browser and the command line. Additional data access examples can still be found on the data lakes page, linked under the main "Data" dropdown.

4. Conclusion

The GDR team is excited to be exploring new technical solutions to old problems, utilizing cutting edge technology and employing agile development methodologies to provide additional features that better connect GDR data to its users and the people of the geothermal community. The GDR continues to serve the community by listening to feedback, remaining agile, and incorporating suggested improvements into future development cycles. These improvements help to better connect geothermal communities to the data they need to inform decisions, advance geothermal research, and develop innovative solutions to local energy problems.

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

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), 17 July 2023. Web. https://gdr.openei.org/.

GitHub. "GitHub." GitHub, Inc., 17 July 2023. Web. https://github.com/.

GTO. "$13 Million in Funding Available for Community Geothermal Heating and Cooling." Geothermal Technologies Office (GTO), Office of Energy Efficiency & Renewable Energy (EERE), U.S. Department of Energy (DOE). 17 July 2023. Web. https://www.energy.gov/eere/geothermal/articles/13-million-funding-available-community-geothermal-heating-and-cooling.

Stack Overflow. "Stack Overflow." Stack Exchange, Inc., 17 July 2023. Web. https://stackoverflow.com/.

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., 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.