Fostering Geothermal Machine Learning Success

Elevating Big Data Accessibility and Automated Data Standardization in the Geothermal Data Repository

Nicole Taverna[1], Jon Weers[1], Scott Mello[1], Adrienne Lowney[1], Amber Mohammad[1], and Sean Porse[2]

[1] National Renewable Energy Laboratory

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

GRC

GEOTHERMAL RISING CONFERENCE

OCTOBER 27-30, 2024 | HILTON WAIKOLOA VILLAGE | WAIKOLOA, HI

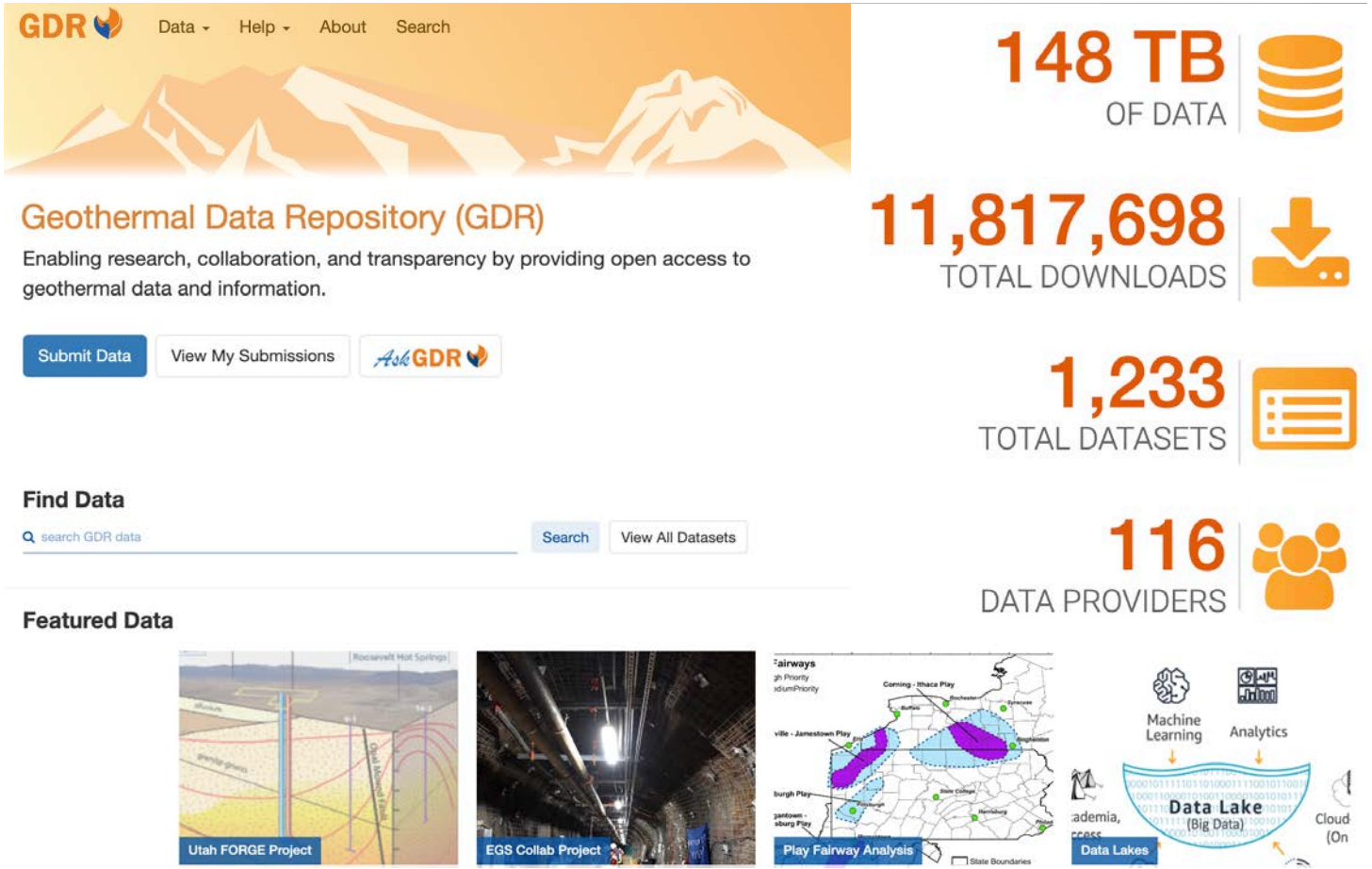
This presentation was produced when the laboratory operated as the National Renewable Energy Laboratory (NREL). The laboratory is now the National Laboratory of the Rockies (NLR).

The Geothermal Data Repository (GDR)

→ Home to data generated by projects funded by the DOE Geothermal Technologies Office

→ Provides public access to geothermal data sets

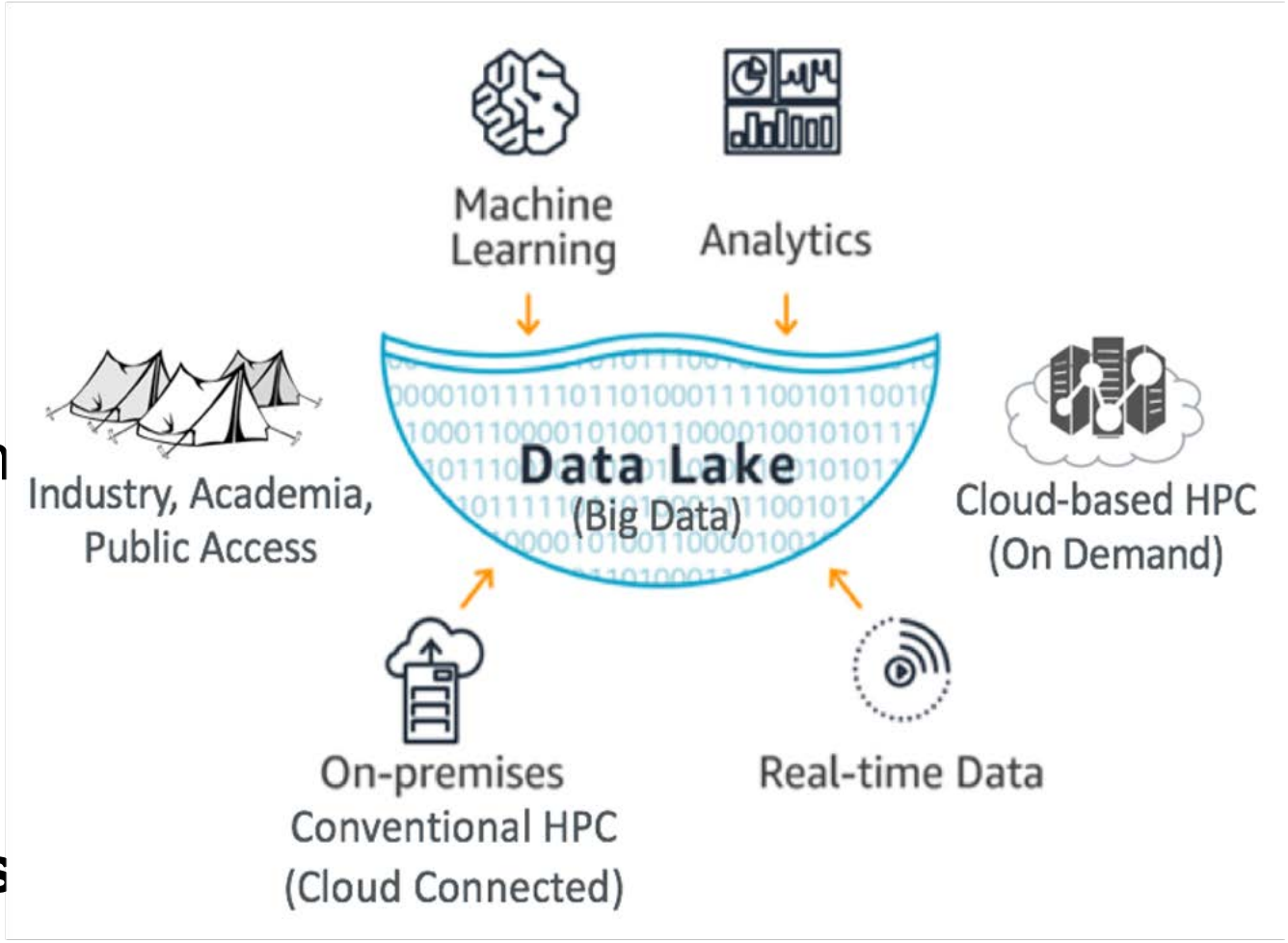
→ **Working to improve the convenience and efficiency of using its datasets in geothermal ML projects**



Read more about the GDR in Weers et al., 2022
Or go to: https://gdr.openei.org/

GDR Data Lakes

- Large or complex datasets (i.e., DAS data) are stored in the GDR data lake
- Allows users to query or work with the data without downloading the full dataset
- Integration with cloud-based tools
- **Improves efficiency in working with large datasets and reduces barriers to access**



Learn more about the GDR Data Lakes: https://gdr.openei.org/data_lakes

Data Standardization and Automated Data Pipelines

Data standardization:

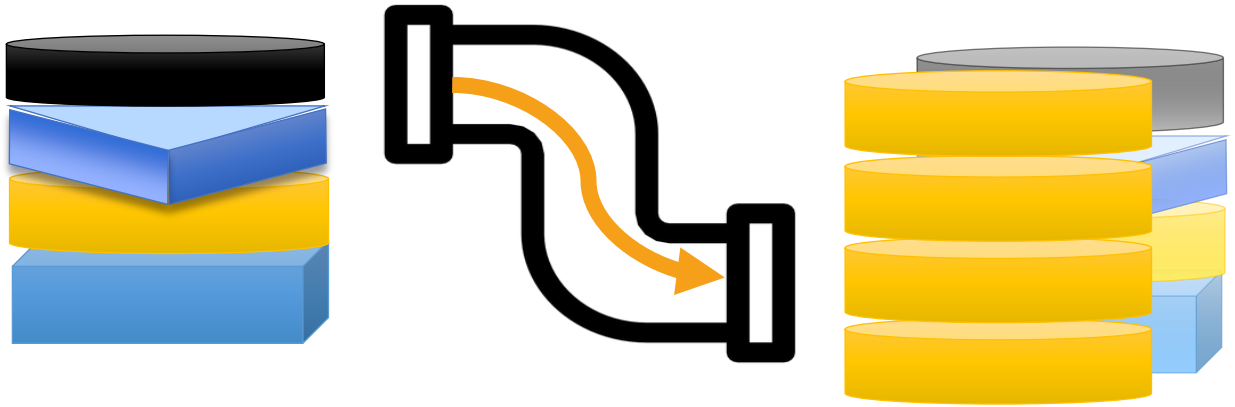
- Improves reusability of data
- Ensures consistency in formatting and content across similar datasets
- Reduces preprocessing requirements & ensures that datasets provide adequate info

The GDR has data standards and pipelines for *drilling data, DAS data, and geospatial data*

Nonstandard data is uploaded to the GDR

Data pipeline automatically detects & translates data into standardized format

Standardized data is added to submission, alongside preserved original nonstandard data



Learn more about the GDR Data Standards and Pipelines: https://gdr.openei.org/standards

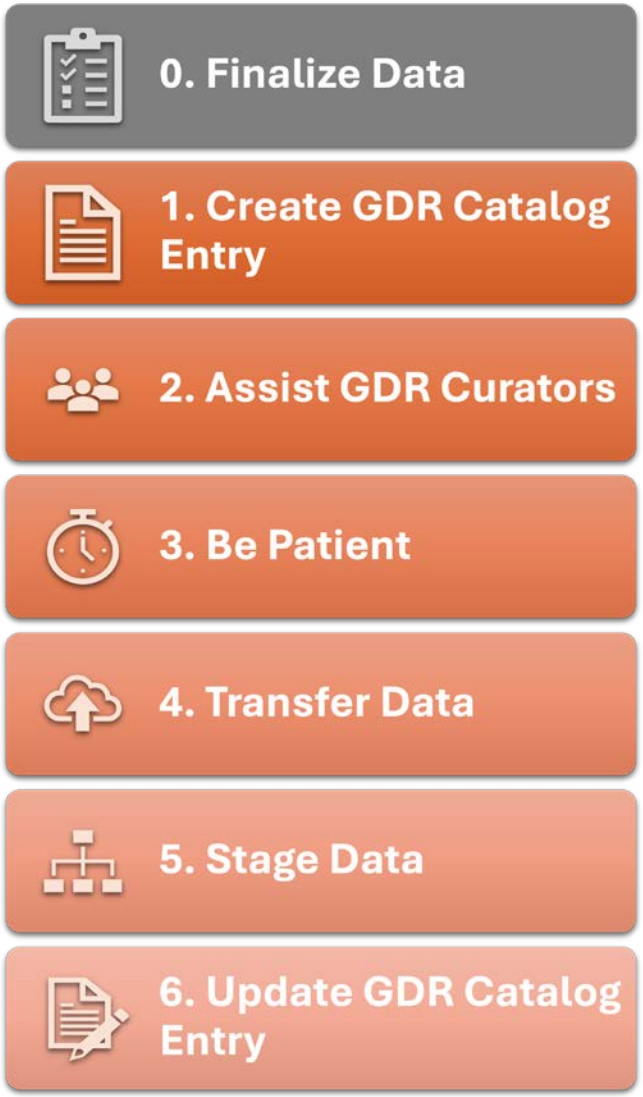
Streamlining the Data Lake Intake Process

Previous inconsistency in data lake submission process:

- Confusing conditional & inconsistent requirements
- Ad hoc communication from curators.
- Lack of clarity on how long the process takes and what it entails.

The GDR has worked to remedy this problem through streamlining the data lake submission process

- Standardizes the user experience
- Makes the intake process more efficient



Direct Access Links for Data Lake Data

- Make info on how to how to efficiently access data from data lakes more readily discoverable
- Allow users to easily copy & paste access commands to access data using the AWS CLI

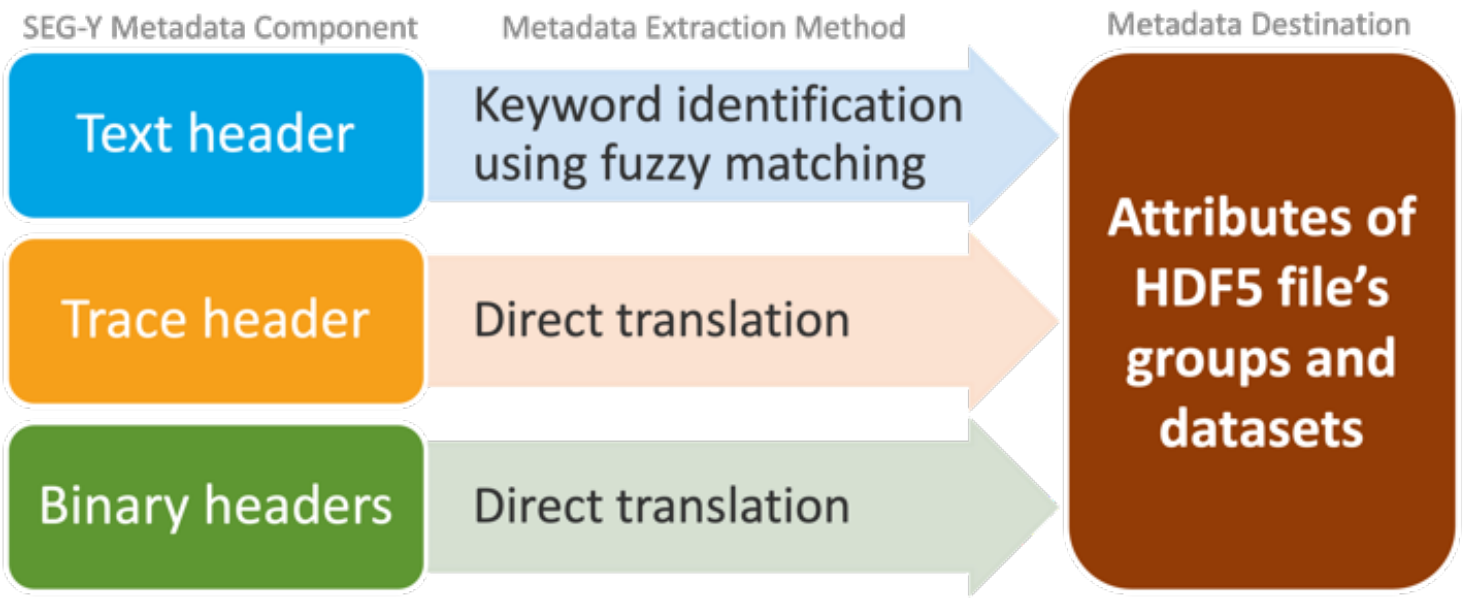
Multiple ways to interact with data in data lakes:

- Jupyter notebooks
- Native cloud query engines like Athena (AWS) and Google Earth Engine
- The GDR Data Lake Viewer
- Native cloud command line tools, such as the AWS CLI example above
- And more!

Learn more about the GDR Data Lakes: https://gdr.openei.org/data_lakes

Implementation of DAS Data Pipeline

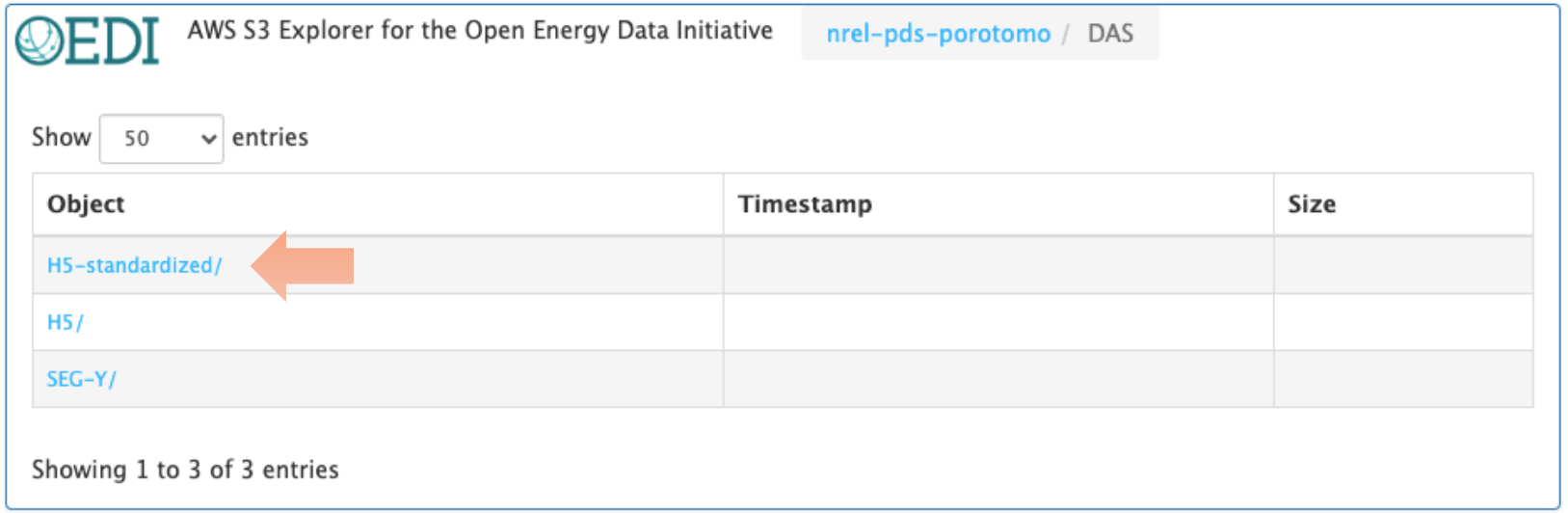
- Built a DAS data pipeline
- Converts DAS data from non-standard SEG-Y format to a standardized HDF5 format
- Based on the DAS RCN's metadata standard and PRODML



Learn more about the GDR's DAS Data Standard: https://gdr.openei.org/das_data_standard

Extension of Data Pipelines for GDR Data Lake

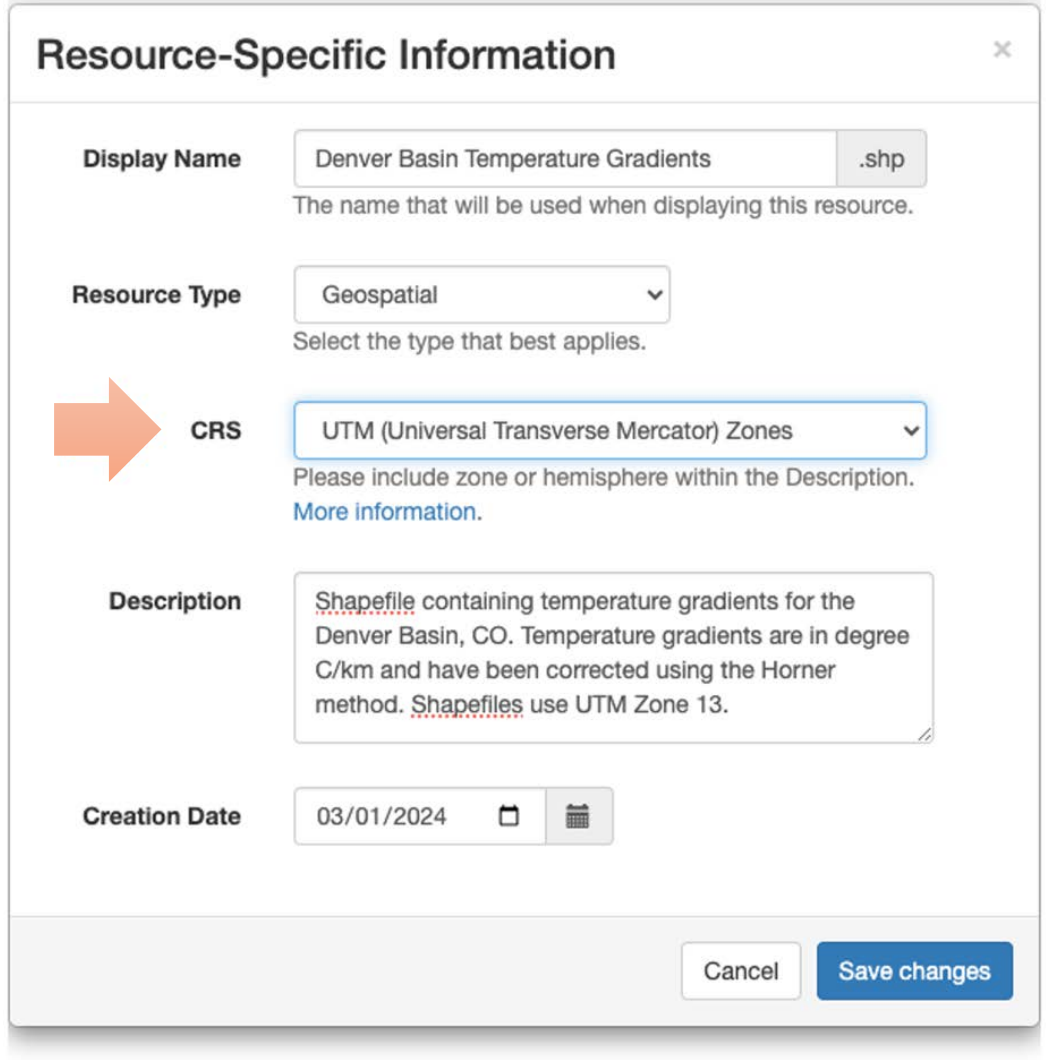
- The DAS data pipeline was adapted to work for data in the GDR data lakes
- Maximizes utility by standardizing the datasets that need it most – *especially large & complex data*



Copyright 2020 Open Energy Data Initiative
Copyright 2014-2018 Amazon.com, Inc. or its affiliates. All Rights Reserved.
Licensed under the Apache License, Version 2.0 (the "License"); you may not use this file except in compliance with the License.
You may obtain a copy of the License at http://www.apache.org/licenses/LICENSE-2.0
This file has been modified by the Open Energy Data Initiative to add links to the OEDI data catalog and add the OEDI logo to provide context, and allow a URL parameter for the bucket name.

New Metadata Requirements for Geospatial Data

- **Added a required metadata field for CRS**
- Spatial domain already required
- Encourages inclusion of other attributes in abstract, resource description, or README file



Learn more about the GDR's Geospatial Metadata Standard:
https://gdr.openei.org/geospatial_data_standard

UX Enhancements

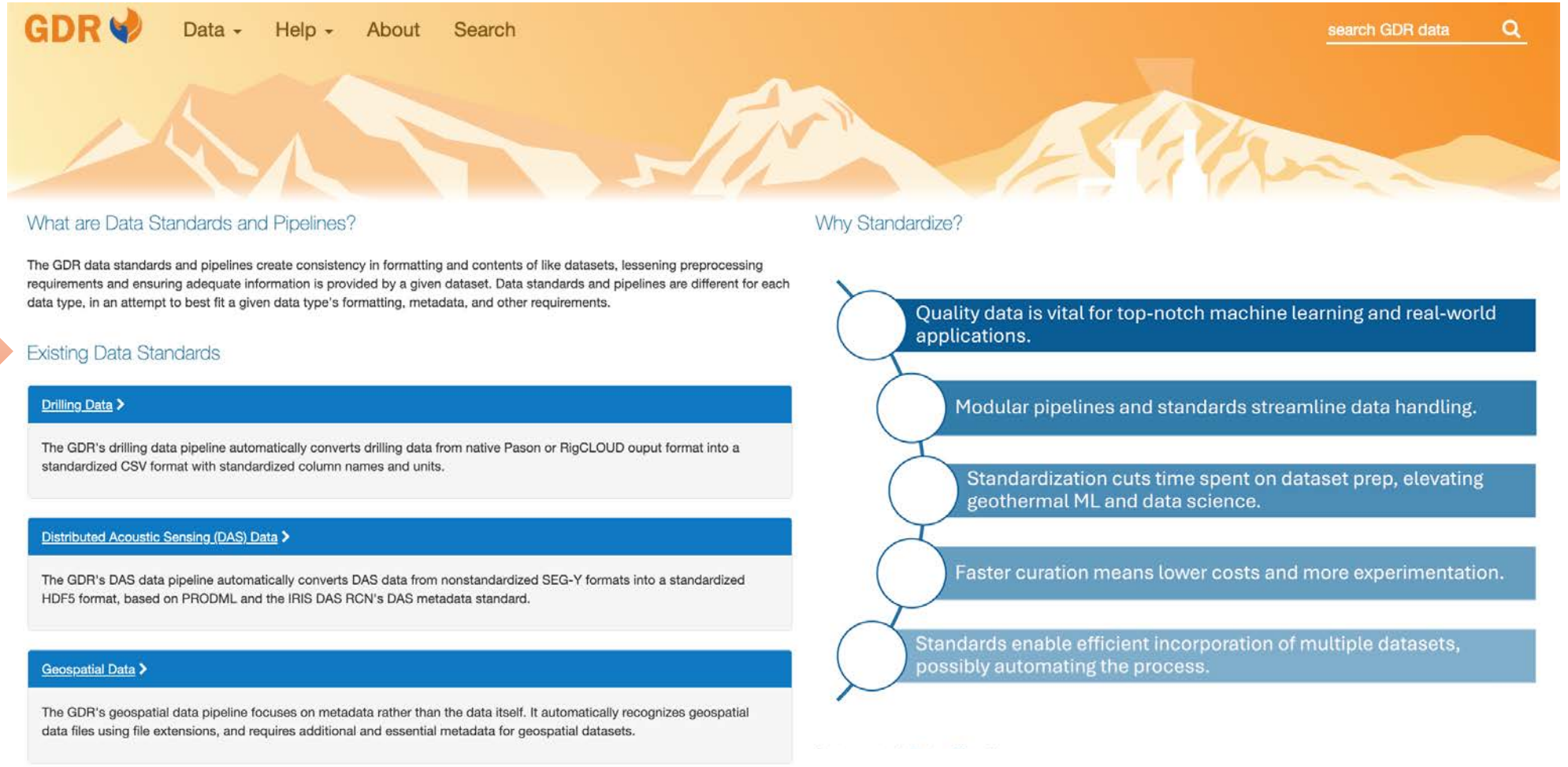
- A UX review was conducted on the existing GDR data standard and pipeline pages.
- This heuristic review primarily focused on the following user-centric design principles:

user needs and goals | **accessibility** | **simplicity and clarity** | **responsive design** | **feedback and performance** | **trust and credibility**

Check out the improved GDR Data Standards and Pipelines pages here:
https://gdr.openei.org/standards

Improvements in Data Standards and Pipelines Pages

Check out our new pages for our existing data standards & pipelines



Check out the improved GDR Data Standards and Pipelines pages here: https://gdr.openei.org/standards

Real-World Applicability

- Too soon to observe impact of standardization
- Top 10 datasets in terms of downloads
  - →6 of these are data lake resources

Rank	Downloads	Year Published on GDR	Title	Dataset storage type
1	5,845,469	2021	PoroTomo Natural Laboratory Horizontal and Vertical Distributed Acoustic Sensing Data	Data lake
2	1,561,471	2021	Brady's Geothermal Field - March 2016 Vibroseis SEG-Y Files and UTM Locations	Data lake
3	1,089,412	2021	Brady's Geothermal Field Nodal Seismometer Data	Data lake
4	728,186	2014	Evaluation Data of a High Temperature COTS (Commercial Off-the-Shelf) Flash Memory Module (TI SM28VLT32) for Use in Geothermal Electronics Packages	Conventional
5	434,950	2021	Brady's Geothermal Field DASH Resampled in Time	Data lake
6	309,362	2023	EGS Collab Experiment 1: Continuous Active-Source Seismic Monitoring (CASSM) Data	Data lake
7	28,441	2014	Project HOTSPOT: Kimama Well Borehole Geophysics Database	Conventional
8	22,472	2023	Imperial Valley Dark Fiber Project Continuous DAS Data	Data lake
9	18,126	2014	Project HOTSPOT: Kimberly Well Borehole Geophysics Database	Conventional
10	12,807	2013	Newberry EGS Demonstration Well 55-29 Stimulation Data	Conventional

**Downloads as of August 26, 2024*

Impact on Research

- Data lake downloads make up most of the FY24 downloads
- The exponential growth in downloads over time would not be possible without the implementation of GDR data lakes.

Conclusions

The GDR has improved its data lakes, data standards, automated data pipelines, and associated pages to educate users

These have reduced barriers related to storage, access, and computation, enabling broader access to large geothermal datasets.

These adaptations have boosted the number of datasets the GDR can accept into its data lakes and made it easier for users to access the data using cloud tools.

Standardizing high-value datasets has improved the quality of training data, thereby enhancing the real-world applicability of geothermal machine learning outcomes.

Automation has increased the availability of standardized data, reducing the burden of project-specific data curation and allowing more time for research.

Overall, these improvements have significantly increased the accessibility and usability of big geothermal data for various projects.

Call to Action

We encourage geothermal data users to:

- Leverage the GDR's new and improved features (data lakes, standardized datasets, and data pipelines)
- Provide feedback on their experience, including both success stories and shortcomings.

Your active participation and feedback are crucial for driving innovation and ensuring that future improvements to the GDR align with your needs and wants

Join us in this initiative to make geothermal data more accessible, efficient, and impactful for the entire community!

GDR

Geothermal Data Repository

Thank You!

https://gdr.openei.org

NLR/PR-6A20-92033

Acknowledgment

This work was authored by NREL 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.

U.S. DEPARTMENT OF ENERGY

NREL

References

- The Geothermal Data Repository (GDR). https://gdr.openei.org/home.
- IRIS DAS Research Coordination Network (RCN) Metadata Working Group. "Distributed Acoustic Sensing (DAS) Metadata Model." Whitepaper (2022). https://github.com/DAS-RCN/DAS_metadata.
- PRODML Work Group, "PRODML Technical Reference Guide." Energistics, v2.2 (2022). https://www.energistics.org/prodml-data-standards.
- Taverna, N., Weers, J., Porse, S., Anderson, A., Frone, F., Holdt, E. 2023. "An Update on the Geothermal Data Repository's Data Standards and Pipelines: Geospatial Data and Distributed Acoustic Sensing Data" *GRC Transactions, Vol. 47* (2023).
- Taverna, N., Weers, J., Huggins, J., Porse, S., Anderson, A., Frone, F., Scavo, R.J. 2023. "Improving the Quality of Geothermal Data Through Data Standards and Pipelines Within the Geothermal Data Repository." *Proceedings of the 48th Workshop on Geothermal Reservoir Engineering*, Stanford University, Stanford, CA (2023).
- Taverna, N., Buster, G., Huggins, J., Rossol, M., Siratovich, P., Weers, J., Blair, A., Siega, C., Mannington, W., Urgel, A., Cen, J., Quinao, J., Watt, R., and Akerley, J. "Data Curation for Machine Learning Applied to Geothermal Power Plant Operational Data for GOOML: Geothermal Operational Optimization with Machine Learning." *Proceedings of the 47th Workshop on Geothermal Reservoir Engineering*, Stanford University, Stanford, CA (2022).

•Trainor-Guitton, W., Martin, E. R., Rodríguez Tribaldos, V., Taverna, N., and Dumont, V. "Distributed sensing and machine learning hone seismic listening." *Eos, 103* (2022).

- Weers, J., Anderson, A., and Taverna, N. "The Geothermal Data Repository: Ten Years of Supporting the Geothermal Industry with Open Access to Geothermal Data." *GRC Transactions, Vol. 46* (2022).
- Weers, J., Porse, S., Huggins, J., Rossol, M., and Taverna, N. "Improving the Accessibility and Usability of Geothermal Information with Data Lakes and Data Pipelines on the Geothermal Data Repository." *GRC Transactions, Vol. 45* (2021).