The Geothermal Data Repository: Five Years of Open Geothermal Data, Benefits to the Community

Preprint

Jon Weers, and Nicole Taverna
National Renewable Energy Laboratory

Arlene Anderson
U.S. Department of Energy

Presented at the 2017 Geothermal Resources Council Annual Meeting
Salt Lake City, Utah
October 1-4, 2017

Suggested Citation
Weers, Jon, Nicole Taverna, and Arlene Anderson. 2017. "The Geothermal Data Repository: Five Years of Open Geothermal Data, Benefits to the Community: Preprint." Golden, CO: National Renewable Energy Laboratory. NREL/CP-6A20-68627. https://www.nrel.gov/docs/fy18osti/68627.pdf

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Conference Paper
NREL/CP-6A20-68627
April 2018

Contract No. DE-AC36-08GO28308

NOTICE

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at SciTech Connect http:/www.osti.gov/scitech

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
OSTI http://www.osti.gov
Phone: 865.576.8401
Fax: 865.576.5728
Email: reports@osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312
NTIS http://www.ntis.gov
Phone: 800.553.6847 or 703.605.6000
Fax: 703.605.6900
Email: orders@ntis.gov

Cover Photos by Dennis Schroeder: (left to right) NREL 26173, NREL 18302, NREL 19758, NREL 29642, NREL 19795.

NREL prints on paper that contains recycled content.

The Geothermal Data Repository: Five Years of Open Geothermal Data, Benefits to the Community

Jon Weers [(a)], Arlene Anderson [(b)] and Nicole Taverna [(a)]
[(a)] National Renewable Energy Laboratory, 15013 Denver West Parkway, Golden, CO 80401-3305
[(b)] U.S. Department of Energy, 1000 Independence Ave. SW, Washington D.C. 20004

Keywords
GDR, geothermal, data, repository, NGDS, information, big data, management, provenance, innovation, cloud, submission, future, NREL, DOE

ABSTRACT

In the five years since its inception, the Department of Energy's (DOE) Geothermal Data Repository (GDR) has grown from the simple idea of storing public data in a centralized location to a valuable tool at the center of the DOE open data movement where it is providing a tangible benefit to the geothermal scientific community. Throughout this time, the GDR project team has been working closely with the community to refine the data submission process, improve the quality of submitted data, and embrace modern proper data management strategies to maximize the value and utility of submitted data. This paper explores some of the motivations behind various improvements to the GDR over the last 5 years, changes in data submission trends, and the ways in which these improvements have helped to drive research, fuel innovation, and accelerate the adoption of geothermal technologies.

1. A tool for the scientific community

The GDR project team is proud to announce support for incremental submissions through a new feature that allows submitters to save a submission in progress and return later, completing it at their convenience. For several years now, this functionality has topped the list of user-requested improvements to the GDR, and we're pleased to say that it is complete and currently available to all submitters. The GDR project team has been working with the geothermal scientific community from day one and has implemented numerous improvements throughout the last five years with the intent of reducing the time needed to complete a submission, encouraging high-quality metadata, and improving the usefulness of submitted data. Enabling users to save a submission in progress addresses several issues affecting data submission and curation through reducing the stress of submitting complex submissions with large or numerous files and submissions with complicated metadata. Submitters no longer need to complete a submission in a single pass. They may now save as often as they like during the submission process, allowing them to preserve the state of their submission after each file upload, change, or other alteration. This also allows submitters of data from group efforts the opportunity to seek more information from collaborators during the submission process without worry of loosing their progress. Lastly, this feature should reduce the overall number of submissions necessary from a single organization, as submitters will now be able to append a submission "in progress", update it over time, and append or remove files at their convenience.

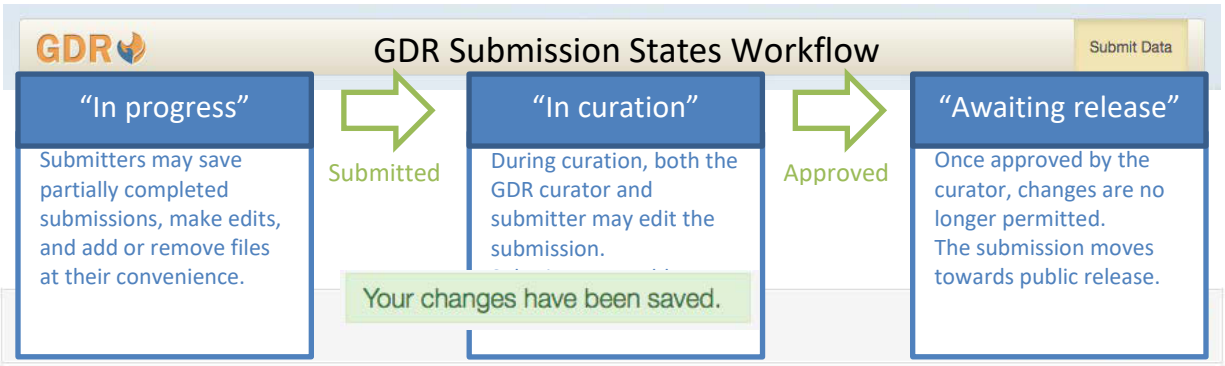

Figure 1: The figure above illustrates the various statuses of GDR data submissions and the actions permitted to submitters during each phase of the submission process.

Additionally, the submission workflow now allows both submitters and curators access to the submission during the curation process. Combined with the improvements mentioned above, this has resulted in a reduction of the effort needed to curate new submissions. Any issues identified by the curator can now be corrected by the submitter while the submission is "in curation", including any issues which may require the removal, addition, or resubmission of a file.

The ability to save a submission in progress is just the latest in a long line of improvements made throughout the history of the GDR. Initially, the GDR supported two additional submission pathways, including an Excel-based template system, in which a metadata file was uploaded along with a single zip file containing the data. The old system also included an option that allowed submitters to provide only the metadata file while physically mailing their data to the GDR project team. These attributes proved problematic as the metadata supplied rarely aligned with the submitted data and often required both parties to do more work to reconcile the differences than required by the current, web-based submission option. Further complicating matters, legal restrictions on the procurement of hardware restricted the media on which data could be physically submitted, occasionally forcing the GDR team to return submissions.

By the second year, the frequency of submissions using one of these legacy options had fallen dramatically, and many experienced submitters developed a preference for the web-based option (GDR 2017). In response to feedback from the community in early 2014, the GDR project team decided to deprecate the other two options and focus development efforts on improving the web-based option. The current submission process is a direct result of user analysis, strategic evaluation of submission quality, and improvements suggested by previous submitters.

Ultimately, the GDR is a tool designed to facilitate communication between members of the greater geothermal scientific community and those performing work funded wholly or in part by the DOE Geothermal Technologies Office (GTO). It has been built from the ground up to communicate data resulting from GTO-funded projects and to increase the exposure of those data to fellow researchers, students, investors, and industry professionals. Initial development efforts were focused on communication with a network of strategic data sharing partners. Rather than endeavor to be the one and only source for geothermal information, the GDR is focused on interoperability, knowledge sharing, and communication of its data catalog with partner sites.

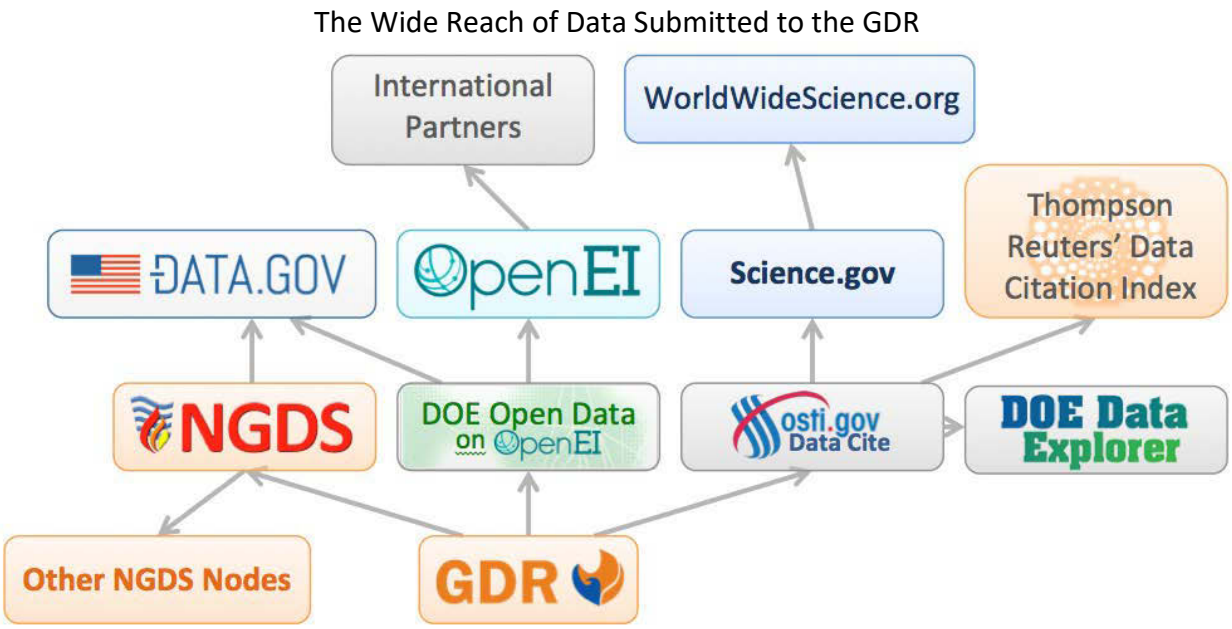

Figure 2: A diagram of the network of geothermal data partners. Data submitted to the GDR is available to users of each of these sites.

In 2013, the GDR was connected to the National Geothermal Data System (NGDS) and the DOE Data Catalog on OpenEI.org, making all data available on the GDR instantly available to users of the other systems as well. In early 2014, the GDR was connected to Data.gov, a federal metadata catalog run by the U.S. General Services Administration (GSA), and to DOE's Office of Science and Technical Information (OSTI), whose system propagates GDR data to their tool, the DOE Data Explorer, as well as the many sites within their own network of data sharing partners, including Science.gov, Thompson Reuters Data Citation Index, and more (Figure 2).

Accurate metrics on the increase in exposure of datasets through a distributed network of this type are difficult to calculate because the metadata passed along by the GDR allows each site to provide their users with the same direct access to GDR data as the GDR itself. An analysis of server log files over several months demonstrates the scale on which the network of partner sites improves the exposure of submitted data.

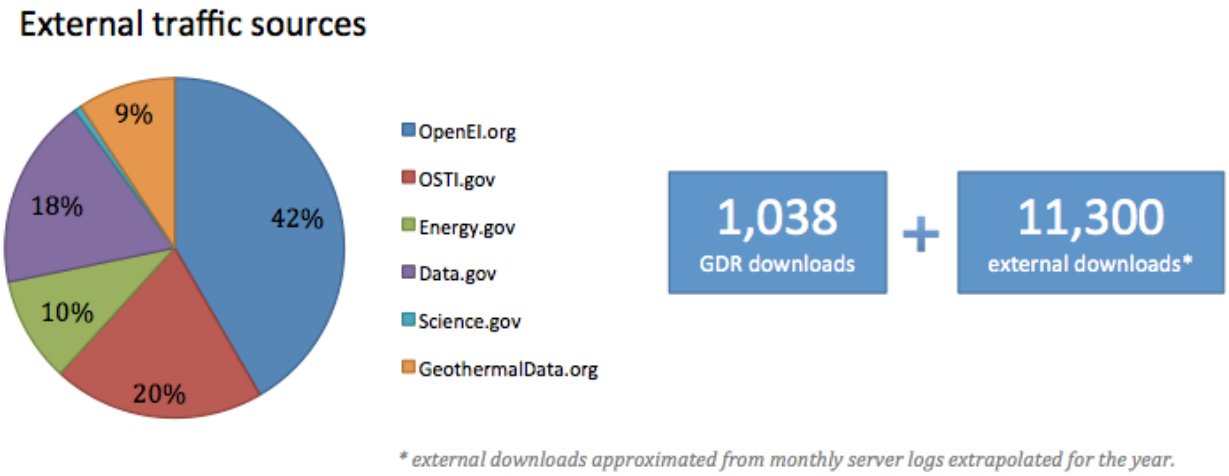

Figure 3: External sources of GDR data downloads from the network of data partners (11,300+) in 2015, compared to direct downloads (1,038) from GDR users.

Data submitted to the GDR then shared with Data.gov, for example, is exposed to an order of magnitude more users seeking information than it would be through the GDR alone (Figure 3). Rather than compete with large, national projects like Data.gov, the GDR has been designed to work with them to maximize the return on value for submitted data. Partner sites like Thompson Reuters (Figure 2) provide an additional avenue for exposure as media outlets occasionally pick up stories about new datasets becoming available and include links to them in their articles. In 2015, an industry news site wrote an article about a GDR data submission titled "Geothermal Drilling and Completions: Petroleum Practices Technology Transfer", a collaboration between the National Renewable Energy Laboratory (NREL) and the Colorado School of Mines (GDR dataset 460: https://gdr.openei.org/submissions/460). The exposure from this article increased the number of downloads of this dataset over the next two months by 300% compared to previous months (Google Analytics 2016).

2. From a simple premise to an open data movement

A report by Deloitt LLP in 2008 identified the creation of a national geothermal data repository as a means of mitigating risk in the strategic adoption of geothermal technologies (Deloitt 2008). In response, DOE funded the development of the NGDS, a distributed network of data repositories all designed to share metadata with one another. The GDR is DOE's node on that network and is the submission point for all data generated from research funded by the GTO.

As an early adopter of open data principals, the GDR was at the forefront of an open data movement that is now a requirement for many government agencies (GSA 2017). Launched in 2010, the GDR laid the groundwork for open data repositories, working with the scientific community to define a system of best practices, and then improve upon them. By leveraging the feedback loop from a close connection to the community, the GDR team was able to quickly incorporate the latest open data strategies, help guide the DOE open data movement, and provide insight to policy regarding the publication of data. By 2011, the DOE Strategic plan included the clause, "DOE's success should be measured not when a project is completed or an experiment concluded, but when scientific and technical information is disseminated" (DOE 2011).

In 2013, the practices already in use by the GDR were institutionalized across departments by a White House Executive Order, asserting that "making information accessible can help fuel entrepreneurship, innovation, and scientific discovery that improves Americans' lives and contributes significantly to job creation" (Burwell 2013).

Data submitted to the GDR is no exception. In addition to a measured increase in exposure, GDR data submitters have enjoyed many of the benefits described by the executive order, including the procurement of additional revenue streams and job creation. On more than one occasion, GDR submitters have been happy to report that the data they submitted at the end of their project lifecycle were discovered by a party willing to fund the continuation of their research. Additional revelry came from analytics collected by the GDR project team. Predominantly, GDR data is used by national labs, universities, government agencies, and industry professionals.

Diving into the details of annual data usage throughout the last five years, the GDR team has identified specific data users (identities omitted for privacy) associated with the development of new geothermal power plants in the US. In some cases, these organizations are directly responsible for the creation of new jobs in the geothermal energy industry. Other organizations using GDR data are tangentially related to job creation by insuring, funding, or otherwise underwriting the development of new geothermal opportunities. These organizations downloaded data from several GDR submissions pertaining to specific areas of interest that were later associated with new development projects (Google Analytics 2016).

3. The value of open data

Opening up access to data is not just exposing them to the public. It also entails opening them up to the possibility of reuse in new and exciting ways. Certain data types and formats are better suited for reuse that others. Submitters to the GDR are encouraged to provide access to raw data along with the summarized, final data products typically associated with their project. This is because summary data is intrinsically biased towards a specific result, while raw data is unbiased and is able to be used in new, unforeseen ways.

The value of raw data has been seen often in the geothermal community, as relatively new technologies and practices like horizontal drilling and engineered geothermal systems (EGS) breathe new life into old core samples, surveys, and other records. This is exemplified in the latest tally of GDR submissions by DOE GTO program area, which shows an augmented number of EGS submissions in recent years (Figure 4).

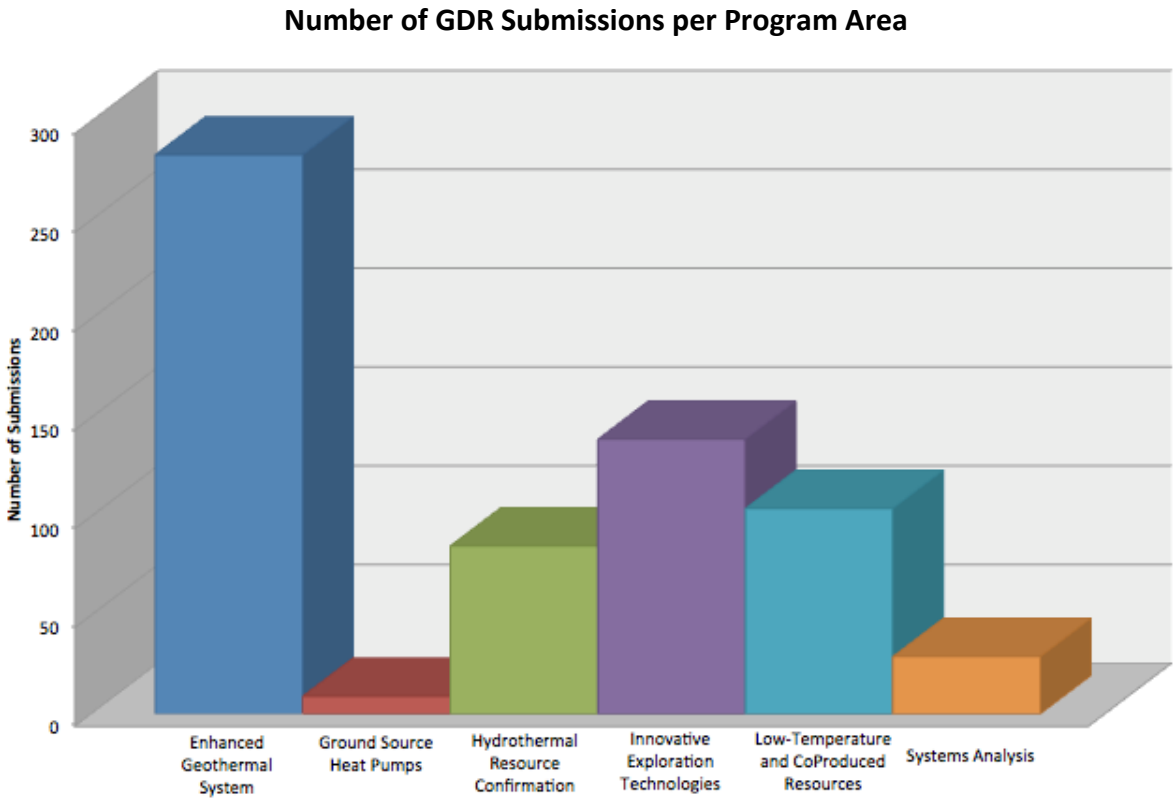

Figure 4: Number of GDR submissions per DOE GTO program area.

Data from sites once dismissed as unfit for conventional hydrothermal are now being re-evaluated for their modern potential. Had only the summary report listing the site's probability for success been available, such re-evaluation would not be possible. Fortunately, many of the raw core samples and other original site data are still available, allowing researchers to evaluate them through a new lens.

Entire projects have been launched on this premise, including DOE's Play Fairway Analysis (PFA) series, the early phases of which applied new research and analysis to existing data sources to determine potential plays for future geothermal projects. The data from many of these analyses are available on the GDR and include comprehensive regional potentials for specific geothermal plays, such as the Cornell University's report on low temperature plays in the Appalachian basin (Figure 5).

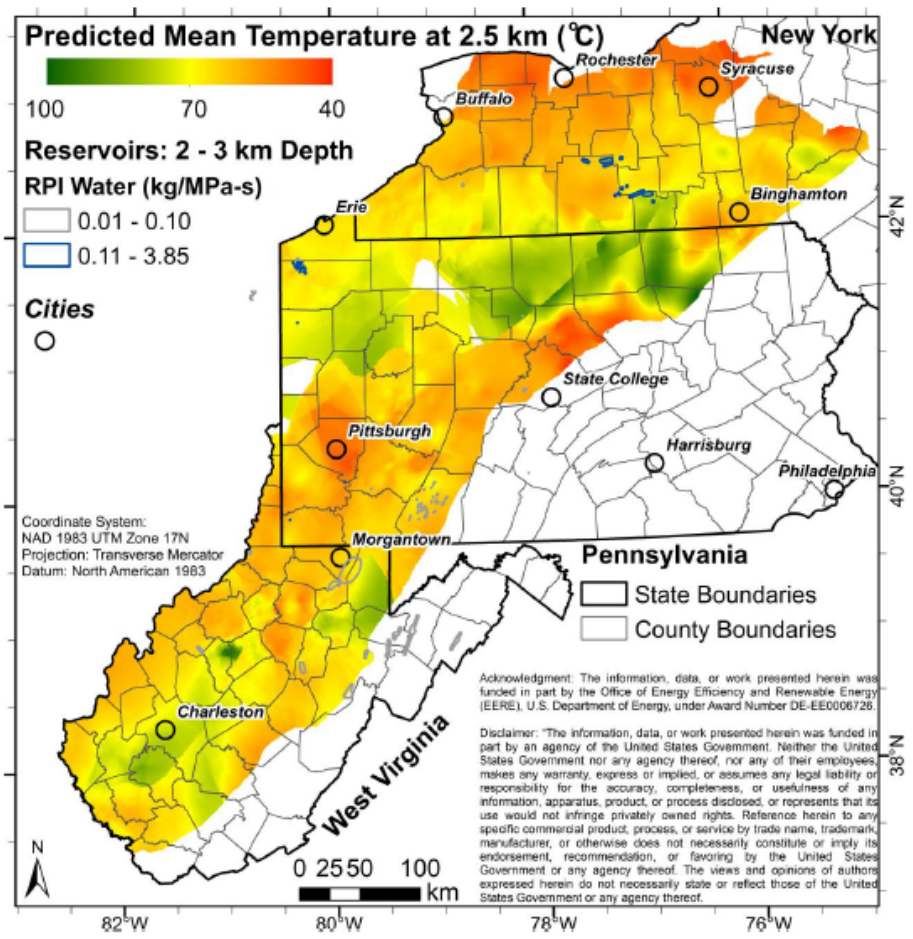

Figure 5: Cornell University. (2017). Final Report: Low Temperature Geothermal Play Fairway Analysis for the Appalachian Basin [data set]. Retrieved from https://gdr.openei.org/submissions/899.

A less obvious use for data is one that spans disciplines. For example, paleontologists have used old geologic core samples from the oil and gas industry to study the fossils of microorganisms at different subterranean levels. Even more impressive, correlations between microfossils present in a sample in pair with the stratigraphy of that sample have given way to new areas of expertise, such as biostratigraphy, which is defined by Brian O'Neill, as "the differentiation of rock units based upon the fossils which they contain" (O'Neill 2017).

The true value of submitting data to an open, publicly accessible repository such as the GDR is reflected in cross-discipline correlations like these and their potential for advancing the advancing new research techniques and the adoption of geothermal technologies.

4. What lies ahead

In the coming years, the amount of data amassed from geothermal research activities will increase dramatically. We are already seeing an increase in the instrumentation of equipment used in labs and production facilities, and the volume of data produced is quickly exceeding our capacity to ingest it at reasonable rates. In 2012, IBM observed that, "every day we create 2.5 quintillion bytes of data – so much that 90% of the data in the world today has been created in the last two years alone" (IBM 2012). Since then, the proliferation of the internet of things,

smart homes, autonomous vehicles, and advancements in instrumentation have dramatically increased the amount of the data generated on a daily basis. Today, autonomous vehicles generate more than 4,000 gigabytes (GB) of data per day, most of it used in real time decision making and discarded just as quickly as it is generated. Only about 12 GB per day (or less than 0.3%) is kept and reported back to manufacturers (Krzanich 2016). Barring a substantial improvement in storage technology, our personal computers and laptops will quickly be outpaced by the volume of data we generate as individuals. With over 270 GB of publicly accessible data, the GDR already contains several submissions that are impractical to store on an average hard drive. Only the top tier of super computers and distributed, cloud-based compute networks will be able to keep pace with the impending tsunami of data. In this future of exponential data generation, curated, quality data will be intrinsically more valuable. Accessibility will begin to include reasonable sizes and processing times, and verbose, accurate metadata will be needed to differentiate specific data from the increasing number of similar datasets.

Additionally, a paradigm shift will likely occur in how data are used for research. As the size of these data begin to exceed the storage capacity of most personal computers, researchers will be compelled to embrace a cloud-based approach to data-driven research in which their hypotheses are transmitted to their data, wherever they may live. Instead of downloading data to a local machine and querying or processing it to acquire results, researchers will need to encapsulate or codify their queries and send them to the data, programming their inquiries to return results to the researcher once processing is complete. This model is already in use in industries that regularly deal with large data formats, and is gaining popularity as an alternative to traditional local storage (Scott 2016).

For the GDR, the future starts with the implementation of additional feature enhancements requested by the geothermal scientific community and continues with the adoption of creative data storage solutions, active data management, and a mission to continue to support researchers, enable innovative uses of data, and accelerate the adoption of geothermal technologies.

REFERENCES

Burwell et al. "Open Data Policy – Managing Information as an Asset." *Memorandum For The Heads of Executive Departments and Agencies, M-13-13*. Director Executive Office of the President, Office of Management and Budget. Washington, DC (2013).

Deloitte LLP. "Geothermal Risk Mitigation Strategies Report." Washington, DC (2008). 28, 41.

Department of Energy (DOE). "Strategic Plan". *DOE/CF-0067*. Washington, DC (2011). 43-44.

GDR. "DOE Geothermal Data Repository." OpenEI: Open Energy Information. National Renewable Energy Laboratory, 25 May. 2017. Web. https://gdr.openei.org.

Google, Inc. "Google Analytics." Google Analytics. Google, Inc., 10 Feb. 2016. Web. http://www.google.com/analytics/.

Google, Inc. "Google Trends." Google Trends. Google, Inc., 12 Feb. 2015. Web. https://www.google.com/trends/explore.

IBM: "What is big data?" IBM: Bringing big data to the enterprise. IBM. 12 Feb. 2012 Web. http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html.

Krzanich, B. "Data is the New Oil in the Future of Automated Driving." *Intel.* Intel Newsroom, 15 Nov. 2016. Web. https://newsroom.intel.com/editorials/krzanich-the-future-of-automated-driving/

Obama, B.: Executive Order, "Making Open and Machine Readable the New Default for Government Information." Office of the Press Secretary, The White House (2013).

O'Neill, B. "Using Microfossils in Petroleum Exploration." University of California Museum of Paleontology. University of California Berkeley. 28 May 2017. Web. http://www.ucmp.berkeley.edu/fosrec/ONeill.html

U.S. General Services Administration (GSA). "Governance." Project Open Data. GSA., 15 May 2017. Web. https://project-open-data.cio.gov/

Scott, W. "Geospatial Platforms: Enabling Disruptive Business Models from Space" *Proceedings: GeoBuiz Summit*, Digital Globe, North Bethesda, MD. 26 April 2016. Web. http://summit.geobuiz.com/presentations/geospatial-platforms-enabling-disruptive-business-models-from-space.pdf

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).