PROCEEDINGS, 41st Workshop on Geothermal Reservoir Engineering
Stanford University, Stanford, California, February 22-24, 2016
SGP-TR-209

The DOE Geothermal Data Repository and the Future of Geothermal Data

Jon Weers [a] and Arlene Anderson [b]

[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

[a]jon.weers@nrel.gov, [b]arlene.anderson@ee.doe.gov

Keywords: GDR, geothermal, data, repository, NGDS, information, node, DOE, big data, provenance, cloud, submission, future, NREL

ABSTRACT

We are, right now, in the midst of the Data Revolution. Just as the Industrial Revolution of the early 1800's marked a major turning point in history with changes to manufacturing that influenced almost every aspect of daily life, the proliferation of big data and the evolution of our collective ability to put it to work for us is having a profound impact on our lives today. The U.S. Department of Energy's (DOE) Geothermal Data Repository (GDR) is leveraging new resources, such as the National Renewable Energy Laboratory's (NREL) new secure cloud environment, and cutting edge approaches to data analysis, to prepare for the challenges that data from energy and geosciences research presents. This paper will explore some of those challenges, highlight some of the benefits of data analysis, and explain the steps DOE and NREL are taking to ensure the GDR is ready for the future of geothermal data.

1. THE DATA REVOLUTION

While some consider the invention of the transistor in 1947 to be the true beginning of the Digital Revolution, it wasn't until much later that the storage of information in a digital format gained momentum. According to a study that analyzed 60 different analog and digital storage technologies between 1986 and 2007, sometime around 1993 digital storage media began to occupy a small but statistically relevant slice of estimated total data storage (Hilbert López 2011). In 2002, the amount of information stored digitally surpassed 50% of all information stored, ushering in the dawn of the Digital Age. By 2007, approximately 94% of all information was being stored digitally, much of it in the form of digital communications, basic profiles, and media such as music, movies and video games (Lindeman and Vastag 2011). Shortly thereafter, smart devices, digital sensors, social media, and the internet of things began to emerge, contributing greatly to data generation. According to IBM, by 2012 humans were creating 2.5 quintillion bytes of data every day, "so much that 90% of the data in the world today has been created in the last two years alone." (IBM 2012).

By 2014, individuals were generating data faster than they could possibly consume it. Every minute of every day, internet users sent over 204 million emails, searched Google 4 million times, and posted over 2.4 million pieces of content on Facebook (DOMO 2014). Meanwhile, massive data collection efforts were being undertaken by grocery and department stores to catalog shoppers' preferences, spending habits, and location (Thau 2014).

Like the industrial and digital revolutions before it, the Data Revolution marks a major turning point in history and has already begun to impact almost every aspect of daily life. During their shared keynote address at a big retail show in 2014, IBM chairwoman and CEO Ginni Rometty said to Macy's CEO Terry Lundgren, "Information is going to be our generation's next natural resource like steam was to the 19th century." (Rometty 2014).

2. THE CHALLENGES OF BIG DATA

For the purposes of this paper, "big data" shall be defined as any set of data too large or too complex to store or analyze utilizing conventional means. The most common example of data that meet this definition are "large file data", or data that consume a cumbersome amount of disk space.

Storing large amounts of data can be prohibitive both financially and practically. Big data storage quickly becomes an exercise in mathematics and probability. For example, a storage solution that costs a mere 2¢ per month per gigabyte would cost 1024 x 1024 x 2¢, or $20,971.52 per month for 1 petabyte. On the hardware side, if the solution utilized an array of 1 terabyte drives, each with only a 0.1% failure rate per year, the solution would need 1,024 drives and could potentially lose approximately 10 drives a year. At these scales, even the most reliable of storage solutions can require additional redundancy to avoid loss of data.

Storage is not the only challenge of working with big data; others include analysis, use, transfer, complexity, and access. Another common example of challenging big data is a data set with a large number of files. A single data set consisting of tens of thousands of smaller files that must either move together or be accessed in a particular order can be difficult to use. In many cases, a special program or other automated solution must be constructed to parse through the files to perform the desired work or otherwise derive meaning

from the data. Even if the aggregate size of the data is manageable, many operating systems limit the number of files that can exist on a drive or within a folder at any given time.

Another example fitting the definition of big data is "time series" or streaming data, and refers to a set of data that are continuously being generated, often from instruments such as a seismograph. While individual slices of data generated from these devices may be small, the never-ending generation of data presents unique challenges to access and utilization. Users of these data must either "tap into the stream" by using the data in live, streaming form, or by preselecting start and end points to "slice" the data into more manageable, finite files.

Data of unconventional complexity are also challenging, and are typically data that capture multiple dimensions of information. They often require special software, powerful computers, or interactive visualization tools to be analyzed. Complex data examples include 3D geologic models or data found in the NREL's Geothermal Prospector (https://maps.nrel.gov/geothermal-prospector), which features multiple layers of geospatial data presented in an interactive analysis platform. The additional work needed to identify relationships between sets of complex data can create big challenges. In the Geothermal Prospector, for example, a geospatial analysis algorithm mines the complex relationships behind different layers and presents results on surface ownership, restrictions, well depth, and temperature ranges for a selected area.

In many cases, the conventional model of downloading data to a personal computer to use it does not work with big data. The personal computer may lack sufficient storage to house the data, the processing power or algorithms needed to parse, query, or analyze it within a reasonable timeframe, or may simply lack the bandwidth needed to transport the data within a reasonable timeframe.

3. ADVANCING GEOTHERMAL INNOVATION THROUGH BIG DATA ANALYSIS

Overcoming the challenges of big data can lead to breakthroughs in innovation. In the retail sector, Kohl's and Macy's are using vast amounts of unstructured data from their customers, including their tweets, emails, pictures, and purchase history, to adjust prices of products in real time based on the latest supply and demand estimates. (FORBES 2014) Target uses big data analyses on abstract purchasing habits across their entire custom base to successfully identify niche markets and target them custom-tailored ads. Most famously, when marketing to expectant mothers, they were able to determine with reasonable certainty that a teen girl was pregnant before her own father knew (Duhig 2012). Their confidence came from subtle shifts in buying habits at key times in their customers' lives, which only became apparent after analyzing the shopping habits of their customer base as a whole. "Many shoppers purchase soap and cotton balls, but when someone suddenly starts buying lots of scent-free soap and extra-big bags of cotton balls, in addition to hand sanitizers and washcloths, it signals they could be getting close to their delivery date." (Duhig 2012) The purchasing trends of over 25 products were involved in this analysis, each individually significant, but together an accurate predictor of pregnancy and even delivery date.

For geothermal, the potential benefits could be huge, especially in the areas of resource exploration. On their own, individual exploration techniques such as seismic reflection surveys, hyperspectral imaging, and gravity surveys can be useful in identifying potential resources. But combined, with data from additional sources and the right analysis, the aggregate results of these individual techniques could produce a model capable of predicting resource potential with far greater accuracy. The oil and gas industry is already employing similar tactics. Shell, for example, pipes data from tens of thousands of seismometers through fiber optic cables to a proprietary server on the Amazon cloud, aggregating their findings from a prospective site in real time and comparing them against findings from thousands of similar sites around the world to paint a more accurate picture of a potential oil reservoir before drilling (Marr 2015).

Of course, Shell benefits from a mature industry and a steady supply of data. This is one reason why submissions to the GDR coupled with proper metadata are paramount to geothermal exploration and development (Weers and Anderson 2015).

4. BUILDING FOR THE FUTURE

Data are most useful when presented in an open, accessible, and widely useable format. The memo accompanying President Obama's May 9, 2013 Executive Order, *Making Open and Machine Readable the New Default for Government Information*, states that "making information resources accessible, discoverable, and usable by the public can help fuel entrepreneurship, innovation, and scientific discovery" (Burwell 2013). In order to encourage big data success like those mentioned above in the geothermal sector, the relevant data must first be accessible. This is a driving force behind the data submission requirements that are often attached to DOE Geothermal Technologies Office (GTO) funded projects. Simply submitting data to an online repository, however, is not sufficient to support big data analyses. The data must also be useful to others. To this end, the GTO has funded the development of official content models (or data models) to encourage the publication of data in industry-vetted formats based on international standards. These models allow data from disparate sources to be formatted consistently, allowing for easy aggregation into complex big data projects. They are available online for anyone to use at http://schemas.usgin.org/models/ (USGIN 2015) and can be accessed directly from the GDR by clicking the "Content Models" link under the "Data" dropdown in the top navigation.

To increase the discoverability of GDR data, developers at NREL and the U.S. Geoscience Information Network (USGIN) are working to improve utilization of data stored in content models to better integrate with the automatic mapping of National Geothermal Data System (NGDS) resources that contain a geospatial component.

Recently, improvements to the GDR have been made to the underlying architecture as well as the curation and submission processes, the most significant of which involves moving to an architecture based entirely in the Amazon cloud. Previously, submissions containing sensitive information were stored on a secure server at NREL. This presented unique challenges for GDR curators, who needed NREL user accounts, separate logins, physical tokens, and a lot of patience in order to curate submitted data.

4.1 Improved Curation

NREL recently obtained an official Authority To Operate (ATO) with sensitive data on the Amazon cloud. The process involved meeting over six hundred specific DOE cyber security requirements and rigorous testing. The resulting approval paved the way for the GDR to migrate to an all-cloud architecture. Currently, sensitive data submitted to the GDR are stored on a special, secure sever within a secure corner of the Amazon cloud. The new architecture permits curators to use their existing GDR account in combination with a soft token to access the data. This solution allows for simple, flexible, faster access to submitted data for curators, and is also more secure.

Figure 1: Illustration of the transition from hard token to soft tokens through the use of Google Authenticator.

Note: This only applies to curators; the login for submitters has not changed.

Simplified access for curators means less time spent logging in and more time spent improving the quality of the metadata associated with submitted data. Additionally, the new architecture grants increased visibility into submitted data, enabling the development of a curation dashboard, which further helps to expedite submission curation and improve metadata quality.

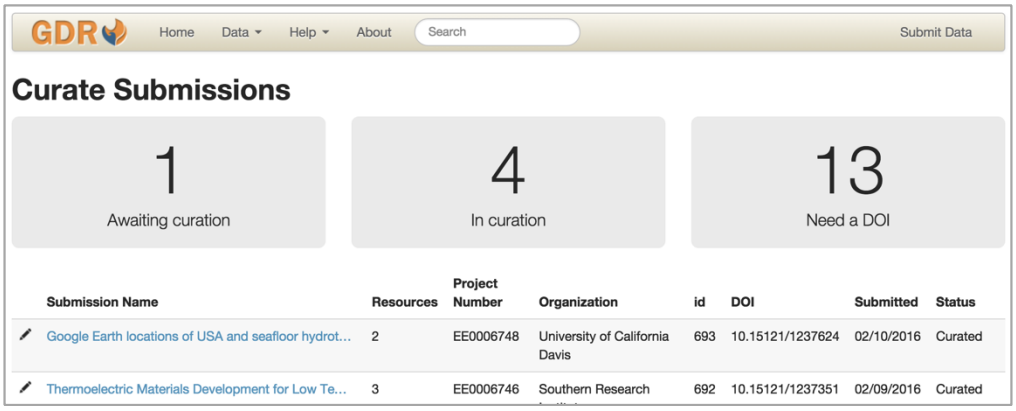

Figure 2: Screenshot of the new Curation Dashboard on the GDR.

The goal of these improvements is to economize the curation process and focus more resources on providing greater quality metadata, which will improve the usability and discoverability of submitted data (Weers and Anderson 2015).

4.2 Scalable Architecture

Now residing entirely on the Amazon cloud, the GDR's scalable architecture is capable of supporting data of all formats and sizes. Limitations still exist, but are now governed by soft guidelines or external factors. One such limitation pertains to submitted file size. Currently, the GDR exhibits a "soft limit" of around 1 gigabyte per file. There is no limit to the number of files a single submission may include and the scalable architecture of the GDR provides a practically unlimited storage capacity, but factors external to the GDR can impose limitations on the size of files that can be uploaded. These limitations can include the submitter's internet connection, their company's internal firewall, or even the number of other programs running at the time of submission.

Another "soft limit" the GDR has encountered in recent years is the cost-benefit analysis DOE must perform on big data submissions. As mentioned above, data of a certain size can become prohibitively expensive to store. Some larger, high-value datasets may be allowed on the GDR, while others may be directed toward alternative submission options.

4.2.1 Options for Big Data Submissions

Depending on the nature of the data, different options exist for big data submissions. The most common types of big data the GDR encounters are:

- Large-sized instrument data, such as seismic
- Data stored in thousands of smaller files, such as daily log files
- Unusually complex data, such as 3D models or geospatial map packages

For large-sized data, the appropriate DOE project leads will evaluate the cost-benefit analysis of storing that data on the GDR. In some cases, a more suitable home for the data may already exist. For example, various universities are host to large, open repositories of seismic data. Any seismic data too large to reasonably submit to the GDR can be stored in one of these repositories and linked to the GDR by using the "Add Link" option in the submission form (shown below).

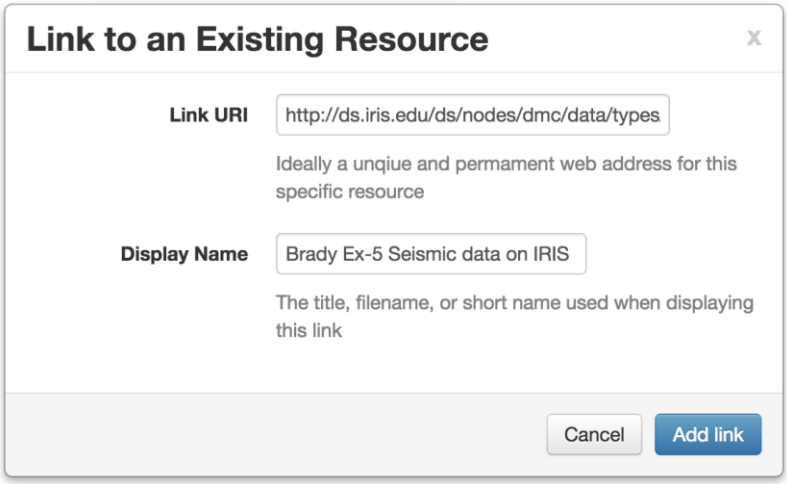

Figure 3: The GDR's "Add Link" form allows the addition of external resources to a GDR submission.

For data too large for conventional download within a reasonable time, a link to a Globus endpoint and access instructions can be added to a GDR submission. Globus is large file transfer management product often used by research institutions.

For data stored in thousands of smaller files, it is recommended that they be aggregated into a single, larger, structured data file, preferably conforming to one of the NGDS content models. If that is not possible, then it is recommended that they be aggregated into more manageable chunks, organized by anticipated use. For example, an instrument generating a 1kB file every minute could easily be aggregated into larger files by day or by month.

Complex data are not necessarily large and can usually be submitted to the GDR with ease. However, these data often come with multiple files that do not function independently of one another. In this case, the individual files are more usable as an aggregate set and it is recommended that they be archived together. Complex geospatial data, for example, are most useful when the individual component files that make up shapefiles are bundled together into a single, cohesive shapefile archives. Each individual geospatial layer should be packed into its own archive to improve discoverability and usability of the data. The use of map packages (.mpk) files is discouraged because the layers within are not independently accessible and also because of the need for a proprietary program to unpack them. Whenever possible, these data should be exported as shapefiles or geo-tiffs before submitting them to the GDR.

The GDR occasionally receives large zip files that are archives of many different types of smaller files. It is recommended that these be unpacked and each type of file be evaluated independently according to the guidelines above. On several occasions, large archives submitted to the GDR have contained numerous extraneous, irrelevant files. In these cases the submissions were rejected and the submitter was asked to resubmit only the relevant files.

4.3 Streamlined Submission

The new GDR architecture has also allowed us to streamline our submission process. In response to user feedback, the GDR has retired the original "3 options" for submission, consolidating them to a single, more convenient option. Clicking the submit button will now take the submitter directly to the submission form, avoiding the confusion of selecting submit options. Submission of archives of files will still be supported, but will now utilize intelligent unpacking algorithms to process them into individual groupings of files, allowing for easy metadata assignment through the web interface. The old "GDR Metadata.xls" file will no longer be utilized. Additional enhancements, including batch assignment of resource metadata will further improve the submission process.

This single-pathway approach to submission will simplify the process, reduce maintenance, and allow the GDR development team to focus their efforts on making submission easier for everyone.

5. BETTER DATA THROUGH SMARTER SUBMISSIONS

Data that are not discoverable are not useful. Several planned, specific changes to the GDR submission process shall improve the discoverability, and therefor the usefulness, of submitted data. These changes include:

- Automatic detection and packaging of shapefile components
- Zip file detection and smart unpacking of contents
- Automatic metadata population based on intelligent assertions from previous submissions, including known user information and detectable file information
- Ability to specify the same location for multiple resources
- Ability to save a submission in progress, make changes, then submit at a later date

All metadata automatically populated by the GDR are meant as a convenience and are fully editable, allowing the submitter to correct any assumptions made that do not accurately reflect the submitted data.

5.1 Metadata Population Guidelines

It is important to think of a GDR submission as a communication with the greater scientific community. Metadata supplied to the GDR are automatically propagated to dozens of external sites, including Data.gov, the NGDS, OpenEI, the DOE Data Explorer, and many more, dramatically increasing the exposure of submitted data:

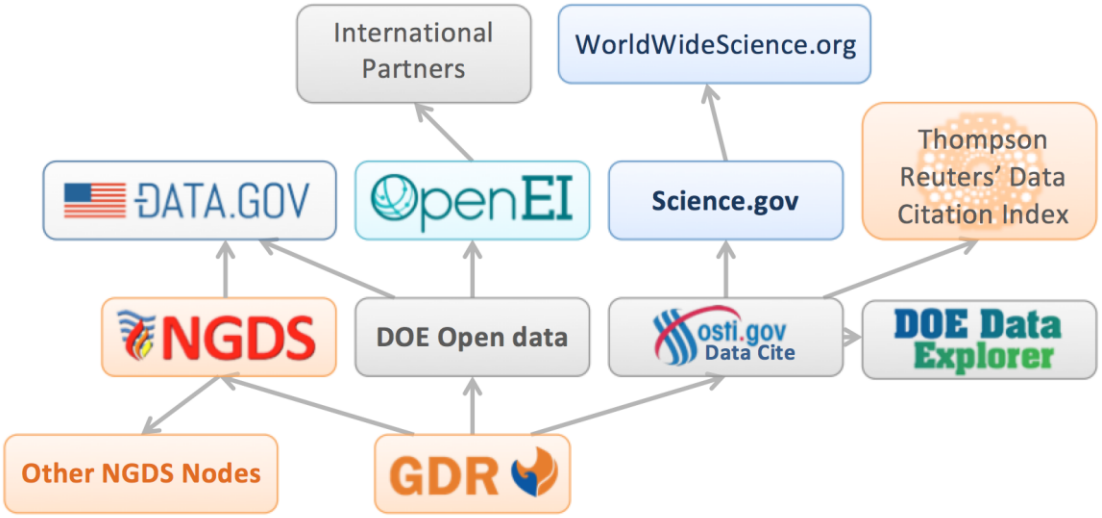

Figure 4: The flow of metadata from the GDR to several of the sites on which GDR data are searchable.

Do not assume domain knowledge or project familiarity when considering the intended audience of a data submission. All metadata fields should be populated with information appropriate for the greater scientific community. For example, the title of a submission should be informative and clearly indicate the nature of the data within. Titles which reference specific project codes, milestones, or contain internal identifiers may confuse potential users of the data. Titles should contain specific information including the subject, location, and nature of the data submitted. Similarly, a good description should summarize the various resources submitted, including where and how they originated, without mention of internal project details such as funding levels, deliverables satisfied, or due dates.

More information can be found on the GDR, in the FAQ section under "Help," and in the document "Guidelines for Provision and Interchange of Geothermal Data Assets", available at http://energy.gov/eere/geothermal/downloads/guidelines-provision-and-interchange-geothermal-data-assets.

6. CONCLUSION

The geothermal industry could benefit greatly from big data analyses, particularly in the fields of predictive maintenance and exploration. Access to relevant data from the oil and gas industry would increase the overall data pool and could amplify the benefits of such analyses. Smarter attribution of metadata can help increase the utility of data generated by the geothermal community, which can help advance research, improve exploration efficiency, and increase the adoption of geothermal energy technologies. Significant improvements to the underlying architecture of the GDR have opened the door for a smarter, more streamlined submission process and

improved curation of submitted metadata. These are just a few of the improvements that DOE and NREL are making to prepare the GDR for the next era of geothermal data.

REFERENCES

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

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

Domo: "Data Never Sleeps 2.0: How Much Data is Generated Every Minute?" Domo. Domo, Inc. 2014. Web. https://www.domo.com/learn/data-never-sleeps-2.

Duhig, Charles.: "How Companies Learn Your Secrets". New York Times Magazine. New York Times. 16 Feb. 2012. Web. http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

GDR: "DOE Geothermal Data Repository." OpenEI: Open Energy Information. National Renewable Energy Laboratory, 15 Jan. 2015. Web. http://gdr.openei.org.

Google, Inc.: "Google Analytics." Google Analytics. Google, Inc., 10 Feb. 2015. Web. http://www.google.com/analytics/.

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

Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. Science, 332(6025), 60 –65. doi:10.1126/science.1200970.

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.

Lindeman, T. and Vastag, B.: "Rise of the digital information age." The Washington Post. The Washington Post, 11 Feb. 2011. Web. http://www.washingtonpost.com/wp-dyn/content/graphic/2011/02/11/GR2011021100614.html.

Marr, Bernard.: "Big Data In Big Oil: How Shell Uses Analytics To Drive Business Success." Forbes Tech. Forbes, Inc. 26 May 2015. Web. http://www.forbes.com/sites/bernardmarr/2015/05/26/big-data-in-big-oil-how-shell-uses-analytics-to-drive-business-success.

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

OpenEI: "Geophysical Exploration Techniques." OpenEI: Open Energy Information. National Renewable Energy Laboratory, 12 Feb. 2015. Web. http://en.openei.org/wiki/Geophysical_Techniques.

USGIN: "Data Exchange Models." United States Geoscience Information Network (USGIN). Arizona Geological Survey (AZGS), 10 Dec. 2014. Web. http://schemas.usgin.org/models/.

Rometty, Gina.: "A New Era of Value: A Conversation with Ginni Rometty." Retail's BIG Show. Jacob K. Jevits Convention Center, New York City, NY. 13 Jan. 2014. Keynote Address.

Thau, Barbara.: "How Big Data Helps Stores Like Macy's And Kohl's Track You Like Never Before". Forbes Retail. Forbes, Inc. 24 Jan. 2014. Web. http://www.forbes.com/sites/barbarathau/2014/01/24/why-the-smart-use-of-big-data-will-transform-the-retail-industry.

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).