DOE Geothermal Data Repository: Getting More Mileage Out of Your Data

Preprint

Jon Weers
National Renewable Energy Laboratory

Arlene Anderson
U.S. Department of Energy

Presented at the Stanford Geothermal Workshop
Stanford, California
January 26–28, 2015

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-6A20-63548
September 2015

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.

DOE Geothermal Data Repository: Getting More Mileage Out of Your Data

Jon Weers [a], 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, USA

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

Keywords: GDR, NGDS, geothermal, data, repository, information, node, DOE, OSTI, Data.gov, metadata, license, submissions

ABSTRACT

All data submitted to the U.S. Department of Energy's Geothermal Data Repository (GDR) is eventually made public. The metadata for these data submissions is searchable in multiple data catalogs, including the GDR catalog and the data catalog on OpenEI.org. Because the GDR is a node on the National Geothermal Data System (NGDS), all data on the GDR are also discoverable through both the Library and the map-based search features of the NGDS. In 2015, NGDS developers will work with the National Renewable Energy Lab (NREL) to further enhance the display and discovery of geospatial data submitted to the GDR directly in the NGDS map search feature, providing an additional means by which data will be discoverable through the NGDS interface. Furthermore, select GDR submissions are also assigned a Digital Object Identifier (DOI), and as a byproduct of this assignment, these submissions are automatically registered in the Office of Science and Technical Information (OSTI) DataCite catalog. From there, these data are federated to additional sites both domestic and international, including Science.gov and WorldWideScience.org. This paper will explore in detail the wide reach of data submitted to the GDR and how this exposure can dramatically increase the utility of submitted data.

1. GETTING YOUR DATA OUT THERE

The submission of data to an online repository is an increasingly required step in many research, development, and exploration activities. Many of these requirements stem from the realization that free, open access to data is beneficial for all parties involved. Now embedded in many U.S. government funding opportunities, these requirements originate from President Obama's May 9, 2013 Executive Order, *Making Open and Machine Readable the New Default for Government Information*, and the accompanying memo from the Executive Office, commonly referred to as the "M-13-13 memo." The memo asserts that "making information resources accessible, discoverable, and usable by the public can help fuel entrepreneurship, innovation, and scientific discovery" (Burwell 2013). These documents specifically require federal agencies to collect and create data in a manner that supports the reuse of the data. Simply making the data available online is insufficient; it must also be useful to others. This illuminates the true intent of the data requirements that come attached to many federal funding opportunities, including those from DOE's Geothermal Technology Office (GTO). DOE's Geothermal Data Repository (GDR), despite predating this memo by almost a year, was created in response to a risk mitigation report from Deloitte LLP that identified the need for a national geothermal database (Deloitte 2008). Developed under similar guidance, the GDR fully supports the intent of the M-13-13 memo and is actively developing new ways to increase the access, exposure, and usefulness of submitted data.

Getting your data out there is not just about uploading it to a data repository, it's about making the data useful to others in way that will foster innovation and promote further scientific discovery. The difference can be subtle. For some, contributing highly reusable data in international standard formats is commonplace; for others, it requires a small shift in thinking and a little preplanning.

1.1 Preparing your data from day one

One of the best ways to ensure that your data will be useful to others is to plan from the beginning to collect and organize your data in a common format. Compiling your data in a widely accepted format, also known as a "standard" or "model", ensures that your data will be easier to interpret and reuse than those that do not conform to a known model. There are a numerous models available online, and selecting the right one can be difficult. Fortunately, the GTO has simplified this task for the geothermal community through the adoption of official "content models" organized for the National Geothermal Data System (NGDS). These industry-vetted models are based on international standards and are available online as empty Microsoft Excel templates at http://schemas.usgin.org/models/ (USGIN 2015). Links to NGDS-supported content models are also available on the GDR under the FAQ section.

Not all data conform to existing NGDS content models. New content models are actively being developed and are open for community discussion at https://github.com/usgin-models. For everything else, usability can be dramatically increased by

planning early to collect several key bits of information and by thinking of the data deliverable as a means of communicating with future users of the data.

1.1.1 Collecting the right information to accurately describe your data

Many data submission portals, including the GDR, have a set of mandatory metadata that must accompany each data submission. Familiarizing yourself with these requirements prior to data collection can make the eventual submission process easier. While a lot of these metadata can be completed at the time of submission, a couple of them – **time** and **location** – are worth taking note of early. Regardless of the nature of the data being created, the collection of precise time and location metadata is critical to increasing the usefulness of an eventual data submission.

1.1.2 Time

Making a note of the time at which the data was created or collected is common, but the precision of the note can often be just as important. Seismograms, for example, are obviously coupled with time. However, rudimentary seismographs might only record time as the number of seconds since the seismograph was turned on.

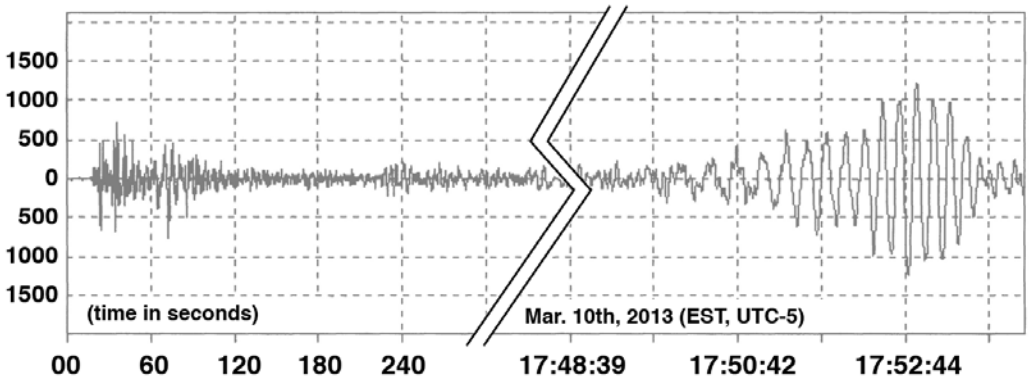

Figure 1: Seismogram showing both poor metadata (left side) and good metadata (right side)

In Figure 1 above, the left side does not provide the exact day and time the seismograph was activated and effectively renders the seismogram useless to anyone not privy to internal project details. Knowing the date and time the seismograph was activated would be helpful, but a researcher hoping to put this seismogram in proper context with other data around the world would need to know the exact date and time, including the time zone or GMT offset. Adding this information can often be trivial, as evidenced in Figure 1 above, on the right side.

Precise time stamps improve the usefulness of all data. Abstract data is no exception. The internet contains numerous groundbreaking studies that produced excellent data but neglected to mention the exact time and date the data were generated. When submitting data on a significant discovery, it can be difficult or even disheartening to consider the future work that could render our data obsolete, but that is exactly what we must do. Without proper time attribution, it can be difficult for researchers to determine which dataset is the most current, especially if a specific generation date doesn't exist for the data. Users of such data often make assumptions about the age of the data using associated dates. For example, a report published this year might actually be the result of analysis of data generated years earlier, and may have produced findings less current than a similar report published last year that used more recent data. For reasons like this, the GDR collects both a "sample date" for each individual data resource and an overall "origination date" for each submission.

For the information in Figure 1 to be more useful, we would also want to know where the seismograph was located.

1.1.3 Location

Today, attribution of location information is more important than ever. Whether you specify it or not, most major internet search engines will append your current location to your search and prioritize your results based on proximity. It doesn't matter if you're searching for "nearby grocery stores" or "geothermal data", search engines such as Google will automatically append your current location to searches that don't specify a location, favoring proximate results (Agrawal & Shanahan 2010). However, studies of search patterns have shown that users of search engines typically include a location in their searches, e.g., "geothermal data Nevada" (Google Analytics 2014). Therefore, including an accurate location in your metadata makes for a more efficient search, increasing the likelihood that searchers looking for your data will be able to find it.

Collecting the precise location is easy when the data are being generated by a piece of equipment or at a physical place, but can be more difficult to define when the data are being generated from a statistical model or multi-regional analysis. In these cases, it is important to define the location in the most specific means possible. To support this, the GDR allows the definition of an area for attributing location.

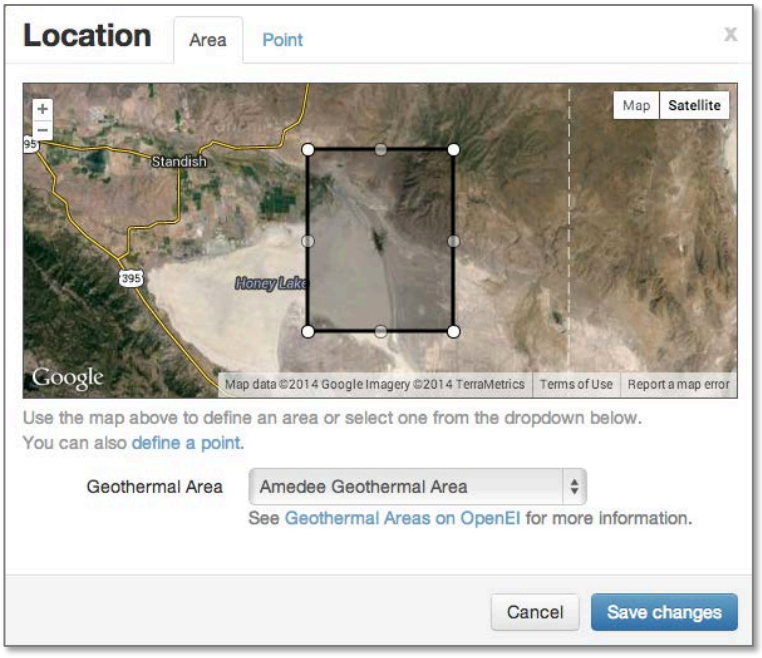

Figure 2: Specifying location metadata in the GDR submission form (GDR 2014, Map from Google Imagery 2014).

Even defining a large area such as the entire United States is better than no area at all. Considering again the example search for "geothermal data Nevada," the code used to power searches on Google, the NGDS, and the GDR recognizes that Nevada is part of the United States and will prioritize domestic results over those in other countries or without specific locations. Even associating your data with the entire world is more useful than data not associated with a location. Doing so would rank your data higher in a search for "global geothermal energy data" than data that were site-specific of unknown location.

1.2 Writing your data to be read

Your data submission is more than a deliverable or milestone. It is a permanent record of your work and a scientific tool intended to enable future research and innovation. Your data should be a self-contained story capable of standing on its own. An easy way to accomplish this is to imagine yourself as a future researcher attempting to use your data for something unrelated to the project from which it originated. Figure 3 below illustrates some of the frustrations created by ambiguous data:

day	t	B flow	B supply T	B return T	Gen	Net	P
1	1:00	4400	148	232	5468	4127	1341
1	2:00	4340	148	232	5420	4078	1342

Figure 3: Example data with ambiguous, abbreviated column headers

In Figure 3, the data contain just enough information for the original creator to differentiate the figures from one another but not enough information for an outside researcher to interpret the data. Assumptions can be made, especially in the first two columns, which appear to be date and time. However, such assumptions can lead to false conclusions. For example, does the second column correspond to hours or minutes? If hours, does the first row correspond to the first hour of work that day or to 1:00am local time? If the latter, what was the local time zone? Appending a data dictionary (an additional file describing the data) to the submission can help demystify the data, but the addition of more descriptive column headers is a more useful solution.

Date Time	Brine flow (gallons per minute)	Brine Supply Temp. (°C)	Brine Return Temp. (°C)	Gross Generation (kW)	Parasitic (kW)	Net Generation (kW)
1-1-2014 1:00:00 am MST	4400	148	232	5468	1341	4127
1-1-2014 2:00:00 am MST	4340	148	232	5420	1342	4078

Figure 4: Example data with well-defined column headers that include units

In Figure 4, the day and time fields have been combined into a single, more precise date-time field. The column headers have been expanded to include full names and units of measure. Finally, the last two columns were reversed to help the narrative flow of net generation as the remainder after parasitic loss is subtracted from gross generation. This last change might appear trivial, but was included because it exemplifies the intent of these changes. The easier a submission is to understand with precision, the more useful the data within it are.

1.3 Making sure it can be used

In addition to making your data useful by including the proper context and required metadata, it is helpful to ensure that people have the legal right to use your data, and that they are aware of that right.

1.3.1 License

With data proliferation across the internet, researchers and other users of data are becoming increasingly sensitive to the usage rights associated with data they find. Making the license clearly visible and easily understood can make your data more useable. Respectable researchers will not use your data if they are unsure of their permission to do so. The GDR automatically attributes the proper license and displays it prominently on the page for each data submission. Additionally, the GDR further encourages the proper use of data from its catalog by providing users with an MLA-style citation (Figure 5), which they can easily copy and paste into their work.

Figure 5: License information and appropriate citation are featured prominently on all GDR submission pages.

2. PROPAGATION THROUGH METADATA DISTRIBUTION

The NGDS is a distributed network of data repositories all feeding into a master metadata catalog. The GDR is the NGDS repository that accepts all DOE GTO-funded data. As part of the NGDS network, all data submitted to the GDR eventually become discoverable throughout the greater NGDS network. Searches for data on the NGDS will include relevant GDR datasets in the results. Clicking on one of the GDR results will provide the user with direct access to the GDR data resource. This is done through the sharing of the GDR's metadata catalog, which conveys the necessary information about all GDR data submissions to enable the remote searching and discovery of those submissions without having to copy the data themselves. The result is an efficient solution where the data is accessible from a variety of locations without excessive storage redundancy.

2.1 Linking Instead of Uploading

Data do not have to be uploaded to the GDR to take advantage of this propagation. The GDR now supports the ability to link to external data resources, meaning any data that are already hosted online can be included in a GDR submission without the need to upload a copy.

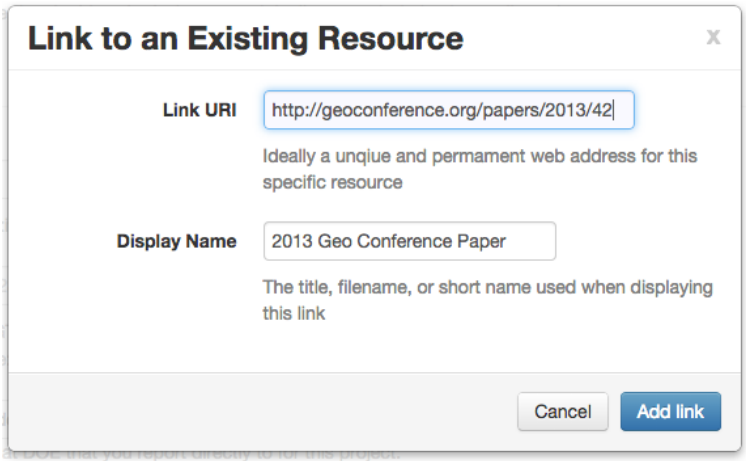

Figure 6: The GDR's "Add Link" form allows the addition of external resources to a GDR submission.

This "link" functionality is ideal for large data, data requiring custom hosting solutions, and data already hosted elsewhere, including copyrighted materials. The GDR does not allow the submission of copyrighted materials, such as proprietary journal articles, but often accepts the underlying data supporting the findings in those articles. In these cases, the underlying data are uploaded to the GDR with a link to the copyrighted article, creating a comprehensive dataset without violating copyright law.

2.2 The far reach of GDR metadata

Available in a machine-readable format, the metadata from GDR submissions can be harvested by any site adhering to one of several international metadata standards that the GDR supports. There is no limit to the number of search portals that could potentially feature GDR data in their results.

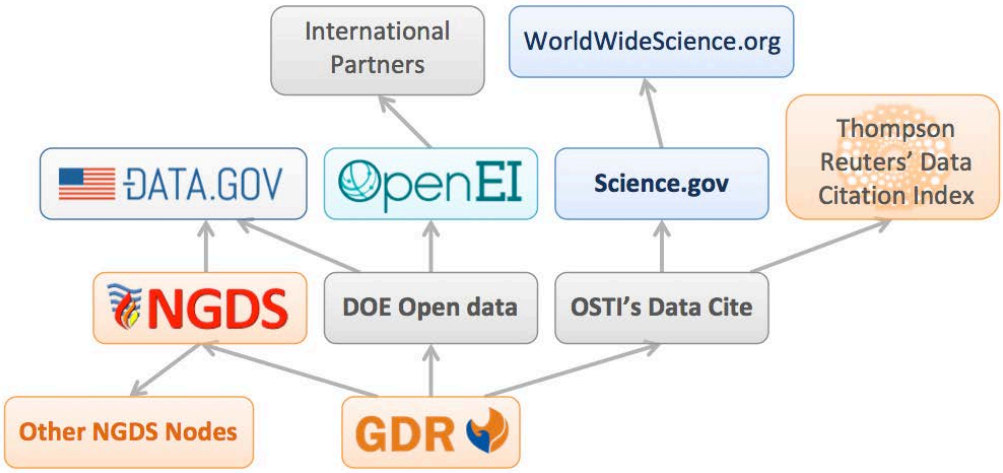

Figure 7: The flow of metadata from the GDR catalog as of January 3, 2015.

In addition to being shared across the NGDS network, GDR metadata also feeds into DOE's open data catalog. From there it federates to the master OpenEI catalog and also to Data.gov, a high-visibility data repository run by the U.S. General Services Administration (GSA). Additionally, as part of Digital Object Identifier (DOI) assignment through the DOE Office of Science and Technical Information (OSTI), select GDR data submissions end up in OSTI's DataCite catalog, which forwards them on to Science.gov, WorldWideScience.org, and Thompson Reuters Data Citation Index.

2.2.1 Expanding the reach

NREL is currently working with NGDS developers to further integrate select GDR data with the NGDS map-based search functionality. Data submitted to the GDR using one of the known NGDS content models will be automatically flagged with additional metadata that will allow the NGDS platform to recognize key location elements within the data itself and render the data directly on the NGDS map-search interface. This additional functionality will take advantage of the superior metadata present in NGDS Content Models to improve the display and discovery of GDR data through the NGDS.

Additionally, DOE is working with external groups to promote both the visibility of available data resources like the GDR as well as the proper citation of data resources in publication. Groups such as the Coalition for Publishing Data in the Earth and Space Sciences, who issued a statement expressing a commitment to "ease transfer of data to repositories" and "promote referencing of data sets" (COPDESS 2015).

3. CONCLUSION

GDR metadata propagates through a network of sites enabling the discovery of GDR data in catalogs around the world. This increased exposure makes the data more readily available to scientists, researchers, and industry professionals who need it, with the goal that more accessible data will help advance research and the adoption of geothermal energy technologies. However, accessibility alone is not enough. The data must also be useful. With forethought, organization, and the inclusion of key metadata such as location, time, clear units, and license, data can be more meaningful and thus more useful to the greater scientific community for years to come.

REFERENCES

Agrawal, R. and Shanahan, J.: Location Disambiguation in Local Searches Using Gradient Boosted Decision Trees, *Industrial Paper*, AT&T Interactive, San Francisco, CA (2010).

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).

COPDESS: "COPDESS Statement of Commitment." Coalition on Publishing Data in the Earth and Space Sciences. National Science Foundation. 14 Jan. 2015. Web. http://www.copdess.org/statement-of-commitment/.

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

GDR: "DOE Geothermal Data Repository." OpenEI: Open Energy Information. National Renewable Energy Laboratory, 15 Jan. 2013. Web. http://gdr.openei.org.

Google, Inc.: "Google Analytics." Google Analytics. Google, Inc, 10 Feb. 2014. Web. http://www.google.com/analytics/.

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

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

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).

Wikipedia Wikimedia Foundation, Inc: "Metadata." 7 Jan. 2013. Web. 15 Jan. 2013 http://en.wikipedia.org/wiki/Metadata.