DOE Geothermal Data Repository – Tethering Data to Information

Preprint

Jon Weers
National Renewable Energy Laboratory

Arlene Anderson
U.S. Department of Energy

To be presented at the Thirty-Ninth Workshop of Geothermal Reservoir Engineering
Stanford, California
February 24-26, 2014

Conference Paper
NREL/CP-6A20-61403
February 2014

Contract No. DE-AC36-08GO28308

NOTICE

This report was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or any agency thereof.

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

Available electronically at 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
phone: 865.576.8401
fax: 865.576.5728
email: mailto:reports@adonis.osti.gov

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

U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161
phone: 800.553.6847
fax: 703.605.6900
email: orders@ntis.fedworld.gov
online ordering: http://www.ntis.gov/help/ordermethods.aspx

Cover Photos: (left to right) photo by Pat Corkery, NREL 16416, photo from SunEdison, NREL 17423, photo by Pat Corkery, NREL 16560, photo by Dennis Schroeder, NREL 17613, photo by Dean Armstrong, NREL 17436, photo by Pat Corkery, NREL 17721.

Printed on paper containing at least 50% wastepaper, including 10% post consumer waste.

DOE Geothermal Data Repository – Tethering Data to Information

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

ABSTRACT

Data are not inherently information. Without context, data are just numbers, figures, names, or points on a line. By assigning context to data, we can validate ideas, form opinions, and generate knowledge. This is an important distinction to information scientists, as we recognize that the context in which we keep our data plays a big part in generating its value. The mechanisms used to assign this context often include their own data, supplemental to the data being described and defining semantic relationships, commonly referred to as metadata.

This paper provides the status of the DOE Geothermal Data Repository (DOE GDR), including recent efforts to tether data submissions to information, discusses the important distinction between data and information, outlines a path to generate useful knowledge from raw data, and details the steps taken in order to become a node on the National Geothermal Data System (NGDS).

1. THE DIFFERENCE BETWEEN DATA AND INFORMATION

To fully understand the difference between data and information, we must explore the process of deriving knowledge from data. In the pursuit of knowledge, we often start with raw data and work our way up. The relationship between data, information, knowledge, and wisdom can be described in a hierarchical arrangement (Ackoff 1989):

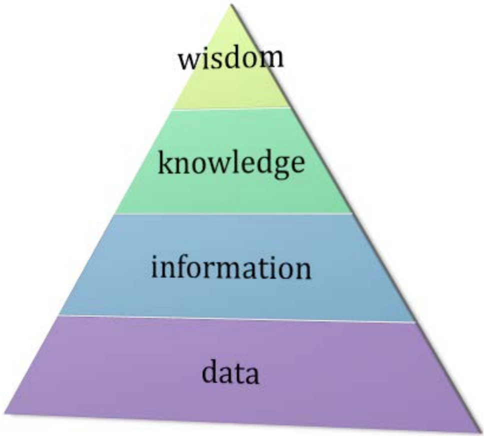


Figure 1: The hierarchical relationship between data, information, knowledge and wisdom.

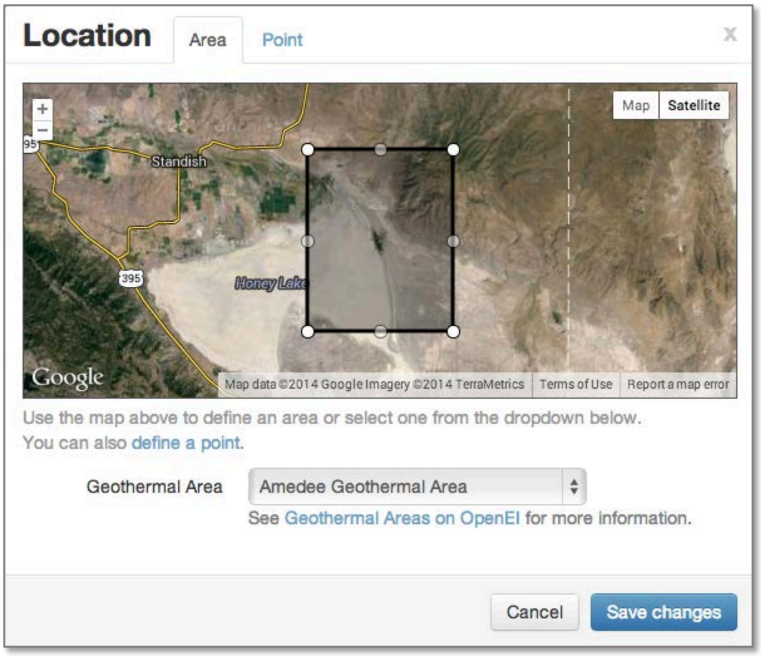
1.1 What are data?

Data start as an observation. In their most basic form, data can be numbers, figures, names, or single bits, which are observed, collected or otherwise assembled in a (hopefully) reusable format. Without context, data are not informative. This is an important distinction to information scientists, as we recognize that the context in which we keep our data plays a big part in generating value. A string of numbers, for example, is difficult to appreciate without context. It is the context associated with data that elevates it to the status of information.

1.2 What is information?

Information is data with context. It is data about something. This contextual association is what gives data its value. The more carefully crafted this association is, the more useful the information becomes. One of the most obvious contexts is the subject of the data. Associating a subject with the data allows us to answer the question, “What is this data about?” The process of creating this association is surprisingly easy to overlook, being so obvious at the time that data collectors often forget to document it, rendering the data unusable by anyone not privy to the conditions of collection.

Another common context is location. By associating data with a given area or locale, we can begin to associate the information generated from the data with other information related to the area. This helps to build context and greatly increases the value of the information. To better enable these associations, the DOE GDR has recently expanded its location metadata to associate individual files within a submission with either an area or a point.



Use the map above to define an area or select one from the dropdown below.
You can also define a point.

Geothermal Area: Amedee Geothermal Area

See Geothermal Areas on OpenEI for more information.

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

Today, attribution of location information is more important than ever. Users of data are searching for information by location in greater frequency than in years past (Google Analytics 2014), and location-aware devices, including modern laptops and mobile phones, are programmatically limiting non-geographic searches to local results using complex algorithms (Agrawal & Shanahan 2010). As a result, data associated with a location become easier to discover and are consequently more useful.

The National Geothermal Data System (NGDS) is a system of interconnected data nodes, of which the GDR is one. The NGDS features a geospatially driven search interface. To empower it, all data federated to the NGDS should include a geographic extent. For these reasons, locations are now mandatory metadata in the GDR. Submissions that apply to entire countries or continents, such as location agnostic technological improvements, should specify those continents or countries as an area in the metadata. Figure 2 demonstrates the area definition during the GDR submission process and is just one of many improvements made to the GDR recently to improve data discovery and interoperability with the NGDS.

Another important context is time. Temporal context can be easy to overlook because present relevance can appear obvious to data collectors. Temporal context is especially relevant in original data submissions. Far too often, original data are submitted without consideration of future iterations on the data. In order to ensure the information is usable for the foreseeable future, the temporal context around its collection must be clarified. For example, the potential generation capacity of a geothermal reservoir might be 50 megawatts. The reservoir may even be commonly referred to as a 50-megawatt reservoir. However, as our knowledge of the reservoir improves, and the efficiencies of our thermal conversion technologies improve, the megawatt rating of the reservoir could change. A more accurate moniker might be “50 megawatts in 2014 using 2012 technology” reservoir. Associating the measurement with the date it was taken and the technology used to collect it can help clarify these types of issues. For this reason, the GDR collects the “sample date” for each resource in a submission and an overall submission “origination date.”

There are a number of other ways to assign context to data. Many of them are, in a sense, pieces of data themselves, which are collectively referred to as “metadata”.

Metadata are commonly defined as “data about data” (Wikimedia Foundation 2013). When properly attributed, metadata serve to define the context of the data they describe, generating information. “By describing the contents and context of data files, the quality of the original data/files is greatly increased” (Wikimedia Foundation 2013).

1.3 Completing the pyramid

Knowledge is understanding. We gather information and apply reason to determine the “why” and “how” of things. The quality of our information is key to our understanding. When little context is available, assumptions must be made to fill the void. Numerous or false assumptions can lead to false knowledge or even unintentional deception. The more context tying our data to information, the more reliable our understanding of that information is.

Our history of knowledge builds wisdom. To use a colloquialism, wisdom comes from experience. From this we can analyze patterns, build predictive models, and anticipate outcomes, none of which would be possible without the proper contextualization of our original data. As the summit of Ackoff's hierarchical pyramid, wisdom is heavily dependent upon the quality of the supporting information, or in our case, the combination of reusable data and descriptive metadata.

2. TETHERING DATA TO INFORMATION

Generating knowledge from data begins with the careful construction of lasting context. To be truly useful, the subject of the data must be indisputable. Ambiguity can lead to doubt and assumption. For example, a statement like "the geothermal potential of Paris is X" can create uncertainty. Those familiar with this paper's origins in the United States might assume the statement refers to Paris, Texas as opposed to Paris, France. Careful attribution of metadata is a reliable way to avoid opportunity for error in reuse.

2.1 Establishing permanence

To be useful time and time again, data must be reliably available. The DOE GDR assigns a Uniform Resource Identifier (URI) for each dataset and resource. A URI is a web address, similar to a URL, but more precise in that it leads to a single, specific (and ideally permanent) identifier for that resource. In the case of the DOE GDR, all resources are given a unique and permanent address derived from their submission number. For example, the following URI will always point to the file *Silver Peak TGH Drilling Map Sep 2009.pdf*, part of submission 268:

http://gdr.openei.org/files/268/Silver%20Peak%20TGH%20Drilling%20Map%20Sep%202009.pdf

2.2 Semantic identity

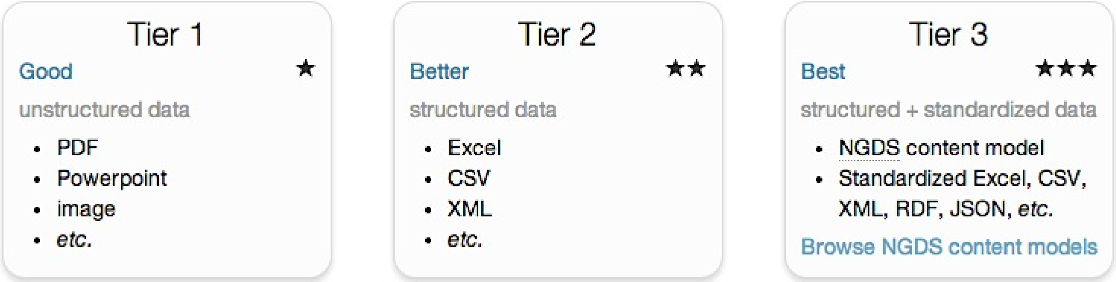
The URI assigned to each resource in the DOE GDR also serves as a permanent reference point for the resource. This identifier appears in the metadata catalog broadcast by the DOE GDR, so that other nodes on the NGDS can positively identify this resource. This is an essential step in linking the data to other relevant information on the NGDS and on any system capable of harvesting the DOE GDR metadata catalog, including OpenEI. OpenEI.org is the open data platform where the DOE GDR lives and it utilizes linked open data to provide worldwide access to energy data.

The DOE GTO is working closely with OSTI and the office of Energy Efficiency and Renewable Energy within DOE (EERE) to link the data submitted by geothermal researchers to program goals and project results. Establishing a semantic identity and is critical to associating key pieces of information with each other, especially when they reside at different locations. To accomplish this, plans for future DOE GDR development include integration with the U.S. Office of Science and Technical Information (OSTI) DataCite platform for the automated generation of Digital Object Identifiers (DOIs). A DOI is commonly used and respected unique identifier for electronic assets (Wikimedia Foundation 2014). Assigning a DOI to a GDR resource would further define its semantic identity and would be linked to its individual metadata and URI. Anyone looking to cite, discuss, or otherwise reference a specific DOE GDR resource could do so without ambiguity by using the appropriate DOI.

2.3 Common ground

Data that conform to commonly used structures or organizational patterns often prove more useful than data that do not. Another important step in tethering data to information is the definition and use of data content models. To date, the NGDS has defined more than 30 data content models (NGDS 2014). The models are a common framework for the organization of data and encourage the submission of data in similar, standardized structures. This allows submitted data to be easily merged, combined, or queried with other data submissions. Storing data in a common format increases the usefulness of the data. As a content model becomes more widely adopted, data conforming to the model become more valuable. Whenever possible, the DOE GDR encourages formatting data in one of the NGDS content models located here: http://geothermaldata.org/content-models/data-interchange-content-models (NGDS 2014).

2.3.1 Three tiers of data



Tier 1

Good ★

unstructured data

- PDF
- Powerpoint
- image
- *etc.*

Tier 2

Better ★★

structured data

- Excel
- CSV
- XML
- *etc.*

Tier 3

Best ★★★

structured + standardized data

- NGDS content model
- Standardized Excel, CSV, XML, RDF, JSON, *etc.*

Browse NGDS content models

Figure 3: The three tiers of data as described on the GDR FAQ page (GDR FAQ 2014).

Data submitted to the DOE GDR that conforms to an NGDS content model is considered to be "Tier 3," or standardized, structured data. Not all submissions will be able to conform to these standards, but those that do will be more reusable (Weers and Anderson 2013).

By contrast, "tier 1" data is unstructured and represents the least useful type of data commonly submitted to the GDR. This tier describes data in formats typically optimized for display or printing. It includes PDFs, presentations, and images. Extracting usable data from an image, especially if that image is a scanned document, can be tedious and prone to error.

The second tier is reserved for structured data. This is typically data in spreadsheets, databases, or structured text formats such as comma-delimited (CSV) files. While the structure may be common, it is often not standardized and will likely require some sort of translation before it can be used with other, similar data. For example, two data submissions might both be structured in spreadsheets, but one describes the location of a resource as "Coordinates" while the other might assign location by "County," mentioning location in a "Description" field, or omit location altogether. Such inconsistencies may make it difficult to compare or combine different data submissions.

Because "tier 3" data submissions adhere to standardized content models, they can be easily combined with other data submissions that use the same model. In a sense, "tier 3" submissions are "plug and play" and can easily be combined with like submissions for cross-cutting analysis or merged into a central database.

3. BECOMING A NODE ON THE NGDS

To become a node on the NGDS, one need only download the NGDS "Node in a Box" software and install it on an appropriate server. The open source software and installation instructions can be found on the NGDS GitHub page: https://github.com/ngds/ckanext-ngds.

For the DOE GDR, this process was complicated by additional security requirements. Fortunately, the NGDS "Node in a Box" software utilizes a common and standard method for data interchange called a Catalog of Services for the Web (CSW), which the DOE GDR development team was able to adopt to marry DOE GDR data to the communication methods required of the "Node in a Box" solution.

The team then resolved discrepancies between the metadata models used by the two sites. Many of these inconsistencies were resolved in the CSW endpoint by translating the DOE GDR metadata from its native format to the NGDS-specific CSW format. In some cases, the DOE GDR metadata identified flaws in the NGDS CSW harvesting tools. In other cases, a change to the DOE GDR metadata collection process was necessary. The location selector featured in Figure 2 is an example of an improvement designed to better marry DOE GDR metadata to NGDS standards.

While numerous improvements have already been made, the DOE GDR team plans to continue to improve the discoverability of GDR data on the NDGS through metadata enhancements and CSW endpoint reconfigurations.

4. TYING IT ALL TOGETHER

The DOE GTO is working closely with OSTI and the office of Energy Efficiency and Renewable Energy within DOE (EERE) to link the data submitted by geothermal researchers to program goals and project results. Tethering GDR data to over-arching goals allows DOE to measure the success of the Geothermal Program as a whole. More importantly, data submitted with proper context is easier to use and easier link with other information, making it more reusable.

> *Linking information from different sources is key for further innovation. If data can be placed in a new context, more and more valuable applications – and therefore knowledge – will be generated. (Bauer and Kaltenböck, 2012)*

Providing the proper context for data is critical to creating these links, which can fuel innovation in the geothermal sector. Careful attribution of metadata can help insure the reusability of the data and advance understanding of geothermal sciences.

REFERENCES

Ackoff, Russell: "From Data to Wisdom". Journal of Applied Systems Analysis 16: 3–9. (1989)

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

Bauer, Florian, and Martin Kaltenböck. Linked Open Data: The Essentials: A Quick Start Guide for Decision Makers. Vienna: Edition Mono, 2012. Print.

"Data Exchange Models." NGDS National Geothermal Data System. US Department of Energy, 10 Feb. 2014. Web. http://geothermaldata.org/content-models/data-interchange-content-models.

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

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

"Frequently Asked Questions." DOE Geothermal Data Repository on OpenEI. National Renewable Energy Laboratory, 15 Jan. 2013. Web. https://gdr.openei.org/faq.

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

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).