PROCEEDINGS, 43rd Workshop on Geothermal Reservoir Engineering
Stanford University, Stanford, California, February 12-14, 2018
SGP-TR-213

The EGS Data Collaboration Platform: Enabling Scientific Discovery

Jon Weers, Bud Johnston, Jay Huggins

National Renewable Energy Laboratory, 15013 Denver West Parkway, Golden, CO 80401-3305

jon.weers@nrel.gov, henry.johnston@nrel.gov, jay.huggins@nrel.gov

Keywords: EGS, Collab, data, collaboration, geothermal, information, management, security, AWS, cloud, future, NREL, DOE, OpenEI

ABSTRACT

Collaboration in the digital age has been stifled in recent years. Reasonable responses to legitimate security concerns have created a virtual landscape of silos and fortified castles incapable of sharing information efficiently. This trend is unfortunately opposed to the geothermal scientific community's migration toward larger, more collaborative projects. To facilitate efficient sharing of information between team members from multiple national labs, universities, and private organizations, the "EGS Collab" team has developed a universally accessible, secure data collaboration platform and has fully integrated it with the U.S. Department of Energy's (DOE) Geothermal Data Repository (GDR) and the National Geothermal Data System (NGDS). This paper will explore some of the challenges of collaboration in the modern digital age, highlight strategies for active data management, and discuss the integration of the EGS Collab data management platform with the GDR to enable scientific discovery through the timely dissemination of information.

1. BARRIERS TO COLLABORATION

In 2017, the number of cyber-attacks and widespread data breaches more than doubled over previous years. (Larson 2017) Cyber security experts are calling for increased protection through implementation of better firewalls and segmentation of networks. (Seals 2017) The deliberate development of barriers between critical digital systems, albeit necessary, is contrary to the effective exchange of information necessary to advance the development of new research projects and the adoption of cutting edge Geothermal technologies. In their 2008 Geothermal Risk Mitigation Strategies Report, Deloitte LLP identified the need for broad access to geothermal data as a means to "reduce the inherent risk in early stages of development and encourage an independent investment market." (Deloitte 2008)

The U.S. Department of Energy's (DOE) Geothermal Data Repository (GDR) was developed in 2012 to provide the greater geothermal scientific community with open access to data and information resulting from work funded wholly or in part by the DOE Geothermal Technologies Office (GTO) (Weers et al 2017). Designed from the ground up to disseminate information, the GDR makes data from GTO-funded projects available on dozens of sites, including Data.gov and the DOE Data Explorer. The GDR also makes these data easily discoverable by users of search engines like Google. The GDR contains a wide variety of data and information from research, exploration, and analysis efforts. Although the GDR does contain select, high-value raw datasets, its primary function is to disseminate the completed, curated results of these activities. The GDR does not provide collaborators with the means of developing interim data work products or organizing large amounts of raw data into meaningful collections.

The amount of data being generated by geothermal research and development activities has increased significantly in the last year. Increases in sensor precision and decreases in installation costs are resulting in more data streams with higher fidelity, while models and simulations are expanding their data footprint in both resolution and complexity. (Weers & Anderson 2016). The sheer volume of data being analyzed in modern research activities can be a barrier to collaboration with some organizations. For example, large file transfer tools like Globus can assist with moving big data from one facility to another, but the recipients must often have a super computer, designated data transfer node, or high-throughput IT infrastructure in order to receive the data. Though commonly available at national labs and larger universities, lack of access to these tools can prohibit collaboration with small businesses or private organizations.

Embattled with increasing cyber threats, information technology experts at national labs, universities, and private organizations are erecting digital barriers designed to restrict access to select data for outside users. These restrictions make the collaborative sharing of knowledge more difficult; a problem which is compounded by the diverse landscape of adopted cyber security strategies. Tools and knowledge sharing methodologies that have been approved for use by one organization are often blacklisted by another. Conflicting cyber security policies can create obstacles to collaboration between organizations where approved knowledge transfer tools are mutually exclusive. For large-scale, collaborative projects teamed with people from numerous organizations, this can be a formidable obstacle to productivity.

Effective collaboration requires all team members, regardless of organization, to have reliable and convenient access to the data and information they need to get the job done.

2. INFORMATION TRANSFER

Early in 2017, the Geothermal Technologies Office awarded a multi-year project, known as "EGS Collab," to a large team led by Lawrence Berkley National Laboratory (LBNL). This team includes numerous participants from other national labs, several universities and private organizations. The purpose of this collaboration is to improve understanding of permeability enhancement and evolution in crystalline rocks at an intermediate scale, which is on the order of tens of meters. A primary goal of these intermediate-scale experiments is progression towards commercial viability of field-scale EGS.

Accurate, timely and objective dissemination of project information will be critical to the achievement of EGS Collab goals. Technologies utilized in this research project include laboratory experiments, borehole instrumentation systems, and THMC simulation codes. Data and information generated by these state-of-the-art technologies will need to be rapidly and effectively shared among project participants and other interested stakeholders.

2.1 EGS Collab Subsurface Characterization and Monitoring Data

The EGS Collab project experimental site is at the Sanford Underground Research Facility (SURF), which is located in the former Homestake gold mine in Lead, South Dakota. Following site selection, the EGS Collab team designed and drilled eight wells from SURF's West Drift on the 4850 level (Fig. 1). Each borehole has been cored and logged and is currently being instrumented in preparation for future stimulation and circulation experiments.

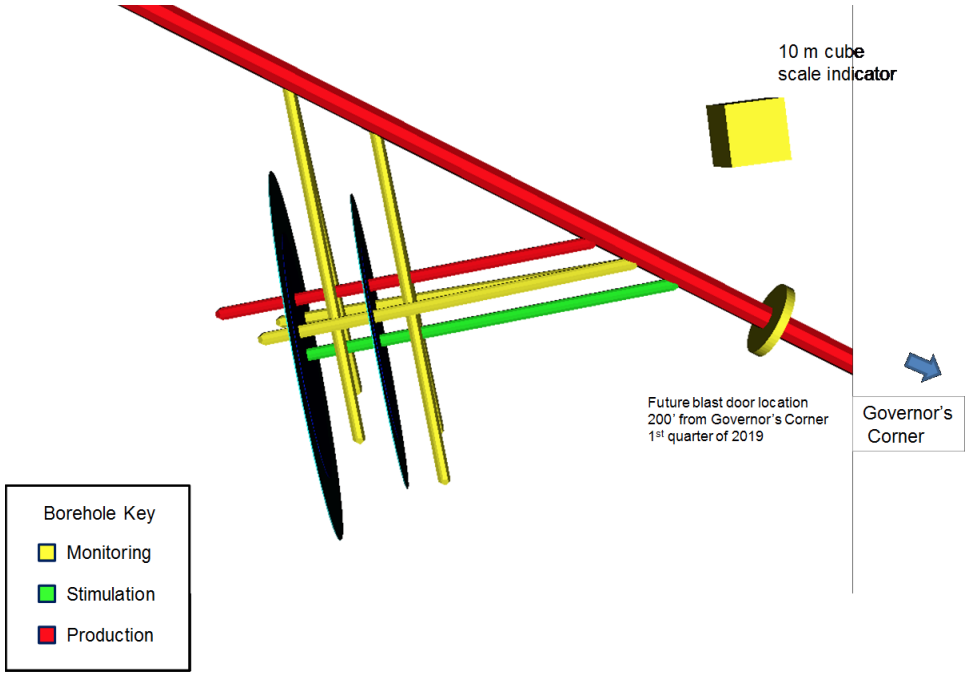

Figure 1 Plan view schematic layout of boreholes along the West Drift on the 4850 level of SURF. Black disks represent potential radial fractures generated through future stimulation experiments. Green borehole represents stimulation well, red borehole represents production well for flow experiments, and yellow boreholes represent monitoring wells. Orientation of stimulation and monitoring boreholes is approximately parallel to Sh-min (Dobson 2017).

Activities are currently underway to characterize the retrieved core and identify foliation, veining, bedding, fractures, and variations in mineralogy (Dobson 2017). Laboratory activities include: (1) baseline characterization of rock samples (e.g., X-ray CT imaging, hydraulic permeability measurements), (2) seismic velocity and attenuation measurements, and (3) core-scale, oriented tensile strength measurements and hydraulic fracturing experiments (Dobson 2017).

Next, boreholes will be characterized with image logs, seismic, electromagnetics, gamma logs, temperature surveys and P- and S- wave characterization across well pairs. Finally, the boreholes will be instrumented for monitoring stimulation and circulation experiments (Fig. 2). These in situ monitoring measurements will include (1) Electrical Resistivity Tomography (ERT), (2) Passive seismic; (3) Continuous Active Source Seismic Monitoring (CASSM); 4) Acoustic emissions and MicroEarthQuakes (MEQs); 5) Distributed fiber optic sensors to monitor changes in temperature (DTS) and strain (DSS); 6) Fracture aperture strain monitoring using the Step-rate Injection Method for Fracture In-situ Properties (SIMFIP) tool; 7) Continuous monitoring of pressure and flow conditions in the injection and production boreholes; 8) Tracer tests; 9) Borehole strain monitoring using tiltmeters; and 10) Wavefield imaging and inversion (Dobson 2017).

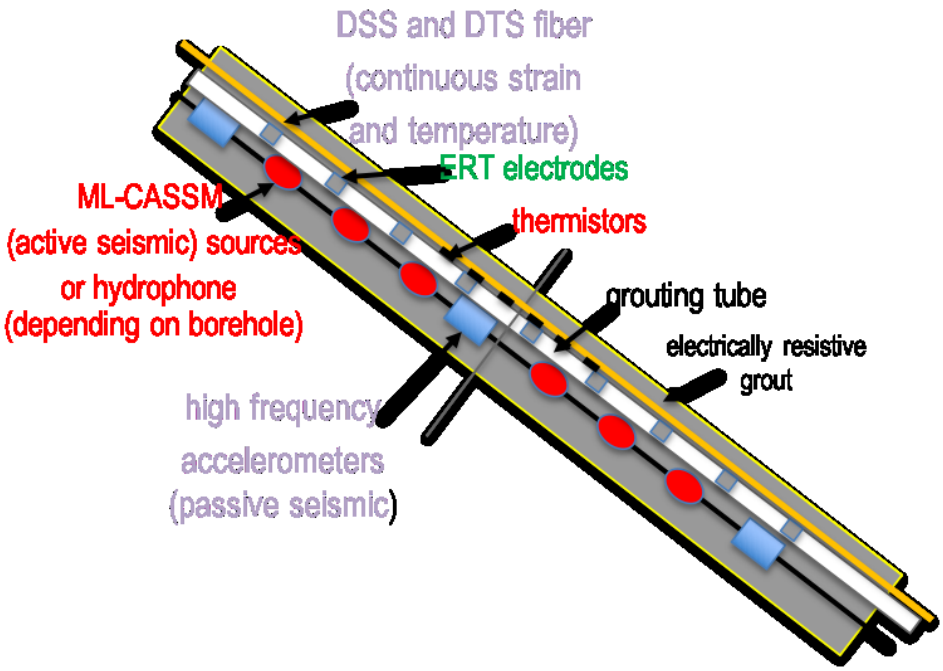

Figure 2 Equipment planned to be installed in 6 monitoring boreholes. Schematic courtesy of Tim Johnson (PNNL), Hunter Knox (SNL), and Jonathan Ajo-Franklin (LBNL).

3. THE NEED FOR A COLLABORATIVE DATA MANAGEMENT SYSTEM

An organized data collaboration system is required for efficient storage, access, analysis and transfer of the abundant subsurface information collected. Technology transfer has been integral early and often in the progress of this research project. Project updates have been well documented in published papers (e.g., GRC, SGW, and ARMA) and presentations (e.g., GTO Peer Review). These interactions have generated interest in the project and valuable feedback has been provided to EGS Collab participants. Now that boreholes have been drilled, cored and logged; information exchange will be critical to the precise development of the fracture planes. Continued data collaboration and regular uploads of data and reports to the GDR will be an essential element to the success of this project.

To facilitate the need for a data exchange and collaboration during this critical phase of the project the Collab data management team worked closely with the GDR development team to construct a Data Management System (DMS) and collaboration space (Collab DMS) that would provide project participants with convenient, universal access to project data while protecting these digital assets and meeting modern DOE cyber security requirements.

3.1 Built for Research and Collaboration

The Collab DMS on OpenEI has been specifically designed to support collaboration in geothermal research projects. Although many online tools already exist that support team-based collaboration and file-based data management, their universally applicable nature falls short of meeting the needs of research-based collaboration. Most notably, their focus on individual users and customizable interfaces is counterproductive to collaborative data organization. Additionally, in order to accommodate a wide range of potential sharing needs, tools like Google Drive or DropBox allow individuals to access select data at virtually any entry point, creating substantial organizational challenges and potential security holes.

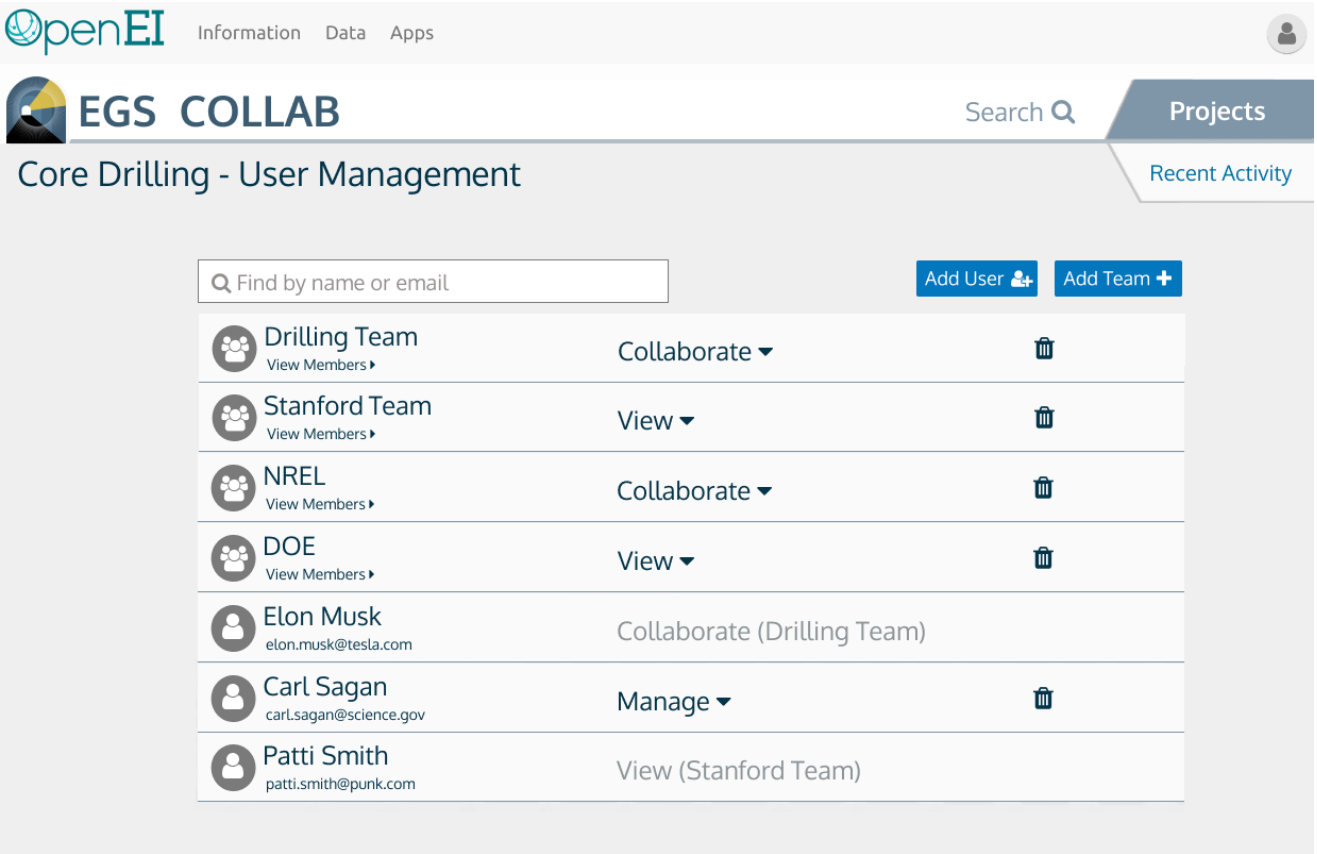

Figure 3 Example Collab DMS interface for managing access to project data for teams and individual users.

The Collab DMS consolidates user access to a single entry point, simplifying team management while presenting all users with a consistent view of the data, which is a prerequisite to better organization of information and proper data management. Team members simply cannot collaborate effectively if they have different views into the same information.

Designed with large projects like EGS Collab in mind, the Collab DMS on OpenEI allows projects to be assigned quickly and easily to known teams as well as individuals. Once created, teams are available to managers of all projects throughout the collaboration space, creating consistency and allowing for easy assignment on future projects. Project managers could, for example, grant a team of DOE reviewers access to view a project with a single click. (Fig. 3)

Project managers can assign large groups of people to a project at once instead of having to assign dozens or even hundreds of individuals. Differing levels of control will allow project managers to grant some groups "view" access while giving others full edit permissions. The user management interface shown in Figure 3 allows teams and individual users to be assigned as either viewers, collaborators, or managers of the project resources.

By focusing development efforts on those that benefit research projects, the Collab development team has been able to prioritize features in the data management system that empower more effective collaboration. For example, users of the collaboration space can easily see the latest changes and additions to a project, or quickly find the last file uploaded from a specific person. These features are possible through the implementation of a detailed history that stores every action a user takes in the collaboration space. The history allows users to quickly see what is new, what has changed, and who changed it, while also providing a detailed audit trail for each resource.

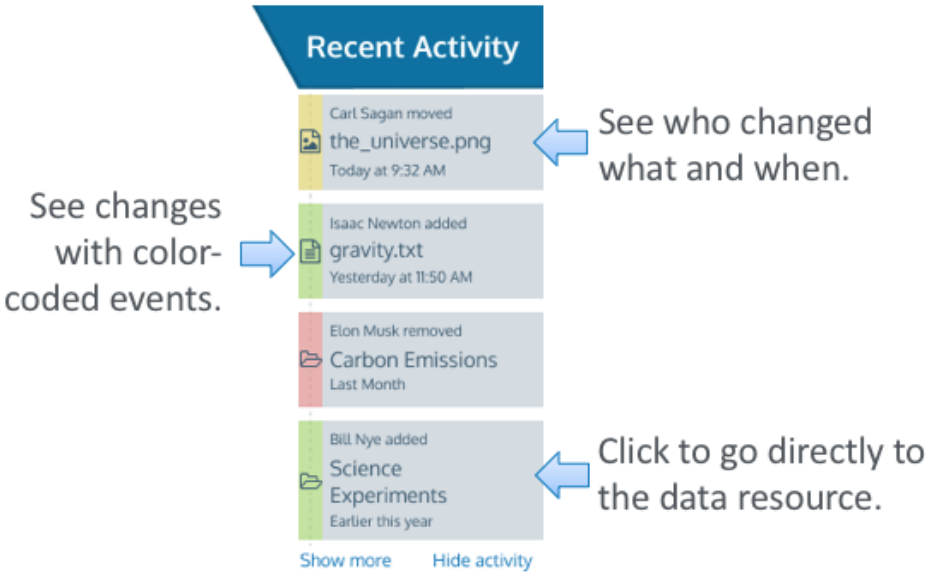

Figure 4 Breakdown of recent activity displayed in the Collab DMS interface.

The Collab DMS has been designed from the start to fill the need for a secure, collaborative environment built specifically to facilitate collaboration for research projects within the DOE Geothermal program. In addition to its focus on research, the Collab DMS was designed to meet current DOE cyber security standards while leveraging existing tools to minimize development costs.

4. REMOVING LIMITATIONS IN A SECURE CLOUD ENVIRONMENT

By extending the OpenEI.org platform and utilizing Amazon Web Services (AWS), the Collab DMS is built on a foundation of existing, proven systems that offer tremendous flexibility and resilience. The Collab DMS shares OpenEI's user-friendly authentication mechanism, and with its AWS backbone supports high bandwidth, nearly limitless storage space, and vast networks that offer redundancy on opposite U.S. coasts for routing, backups, and system security. The cloud-based environment has been vetted and approved by DOE cyber security.

4.1 Universal Secure Access

OpenEI's authentication system is not tied to any single lab or other organization's internal user management system. Instead, it employs the latest DOE-compliant security measures to provide a simple, convenient user registration process. Verified users of the Collab DMS may log on with a single click using an existing Google account or by providing an email address and password. The secure, shared authentication mechanism allows users to navigate seamlessly across other OpenEI properties such as the GDR and OpenEI datasets without having to sign in again. Privileged accounts are protected by a secure yet user-friendly two-factor authentication mechanism, where the trusted device can be the user's own smart phone.

4.2 No File Size Limits

The Collab DMS supports nearly unlimited file sizes, effectively limited only by the reliability of the connecting network during transfer. The Collab DMS file uploader breaks larger files into manageable 10Mb chunks and can upload as many of these as needed, provided there is a network connection, making it possible to transfer files of 1Tb or more. The Collab DMS can store any number of files of this size because it utilizes the Amazon Elastic File System (Amazon EFS), which is "elastic" in the sense that it expands automatically to meet demand, providing petabytes of potential storage, much like Amazon S3, but with response times similar to a traditional hard drive. Amazon EFS is also highly reliable with automatic replication to multiple physical locations. For security, pre-release data is stored encrypted on the EFS system behind multiple firewalls.

4.3 DOE Approved and Responsive

Unlike general purpose public file sharing sites, the Collab DMS was designed to meet stringent Federal Risk and Authorization Management Program (FedRAMP) requirements, making file sharing accessible for users under any branch of DOE. Additionally, it adheres to cutting-edge web standards featuring a modular code base capable of quickly adapting to changing user and security needs.

4.4 The Secure Cloud

The Collab DMS utilizes AWS to ensure fast, secure, robust and reliable networks, hosting and data storage. With an infrastructure spanning continents, the AWS secure cloud promises minimal latency and the most reliable uptime possible with vast distributed networks,

and critical data storage automatically duplicated in isolated data centers in multiple geographic areas (termed Availability Zones) across the country. In practice, if one data center loses power temporarily, or in theory, if it should be destroyed by a natural disaster such as a flood, traffic is automatically rerouted to another data center with redundant versions of the lost servers and copies of all their data. And for security, sensitive data is encrypted from the client's computer, and protected in transport using Secure Socket Layers (SSL) to secure AWS data centers, encrypted within AWS with Transport Layer Security to a firewalled, restricted network known as an Amazon Virtual Private Cloud (VPC), where it is stored in an encrypted virtual file system. Industry leading security teams at AWS monitor vulnerability reports and can implement multilayer mitigation solutions from Distributed Denial of Service (DDoS) attack prevention to automatic security patching, and the Collab DMS team monitors Nagios alerts to identify and mitigate situations before they become problems. Highly granular account administration for system administrators is controlled by AWS Identity and Access Management (IAM) rules and two-factor authentication, and transactional email deliverability is ensured by the reputation of OpenEI.org and the reliability of the AWS Simple Email Service.

The Collab DMS incorporates conveniences and an infrastructure that could only have been imagined a few years ago and enjoys these technologies with minimal development costs by integrating with the existing OpenEI.org architecture, capitalizing on the team's architectural and development expertise, and using Amazon Web Services, not just as a hosting provider, but for its built-in flexibility and resilience.

5. CONNECTING TO THE COMMUNITY

The Collab DMS is built on the same underlying technology as the GDR, allowing for quick and seamless transition from the collaborative working environment to a finished, completed data submission. This allows the Collab DMS to leverage the GDR's vast information dissemination capabilities, providing a clear pathway for researchers to disseminate their findings to the greater geothermal scientific community.

The Collab DMS is designed to be a working space for data and information to be coalesced, cleansed, and transformed by collaborators from multiple organizations. It is anticipated that only a portion of the data that pass through this tool will be suitable for eventual publication and included to a finished data product. The Collab DMS allows users to organize their data and, when ready, select a portion of it to send to the GDR.

Figure 5 The Collab DMS allows any folder to be published to the GDR.

5.1 Integration with the DOE Geothermal Data Repository

Users of the Collab DMS can select any number of individual folders and send them to the GDR as data submissions. (Fig. 5) A wizard built into the Collab DMS helps the user organize the contents within the folder to meet the specific submission requirements of the GDR. Metadata automatically collected from the Collab DMS, including metadata on authors, data resources, and the associated project will be sent with the data to the GDR to automatically populate as much of the GDR submission form as possible. The partially completed form will open in a new browser window to allow the user to quickly and conveniently complete the submission from within the existing GDR interface.

Development for the Collab DMS has been streamlined by tightly coupling it with the GDR, leveraging the existing GDR infrastructure for data storage and user authentication, and utilizing the existing GDR submission workflow to publish EGS Collab data to the greater scientific community. Data from the Collab DMS that have been sent to the GDR will arrive as a new submission and will go through the regular GDR curation and dissemination process. Like any GDR submission, these data will have the option of being released as soon as they have been curated, or at some future date. Once publicly available, these data will be disseminated to all of the existing GDR data partners, including Data.gov, the Office of Science and Technical Information's (OSTI) DOE Data Explorer, and Thompson Reuters' Data Citation Index, and more. (Fig. 6)

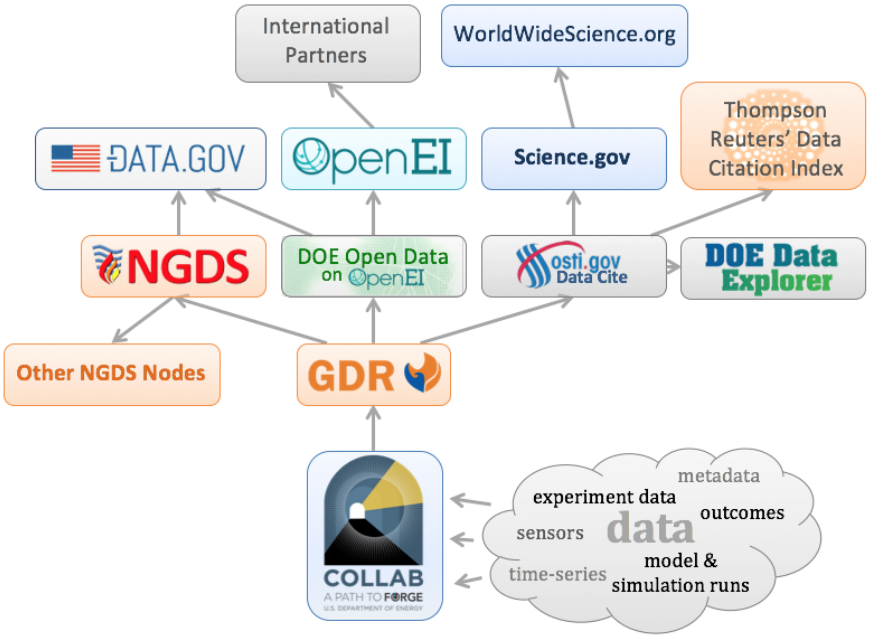

Figure 6 Data propagation through the network of GDR data partners available to EGS Collab.

The closely integrated design strategy implemented for the Collab DMS significantly reduced the cost of development for the platform and avoided any duplication of effort while preserving existing workflows for data submission. Leveraging existing workflows to generate quality, curated data avoids requiring researchers to learn an additional system and allows them to focus on their research.

6. EMPOWERING RESEARCH

The primary goal of the EGS Collab data management platform is to empower researchers on large collaborative projects to quickly and easily share data and information, to reduce barriers to collaboration and to streamline the development of quality, curated data products. Whether sharing information during an experiment on the Collab DMS or publishing the results to the world through its close integration with the GDR, these tools reduce obstacles to research and enable scientific discovery through the timely dissemination of information.

6.1 Future Potential

The Collab DMS has been developed using a modern, agile methodology and modular components that allow it to easily be adopted for other purposes. The open, flexible architecture employed by the platform reduces the need for custom data management development and allows it to easily be adapted for use by future projects.

ACKNOWLEDGEMENT

This research was supported by the U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy (EERE), Geothermal Technologies Office (GTO) under Contract No.DE-AC36-08-GO28308 with the National Renewable Energy Laboratory as part of the EGS Collab project.

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.

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.

Dobson et al: An Introduction to the EGS Collab Project. *GRC Transactions Vol 41*, 41st Geothermal Resources Council Annual Meeting, Salk Lake City, UT (2017).

GDR: "DOE Geothermal Data Repository." OpenEI: Open Energy Information. National Renewable Energy Laboratory, 15 Jan. 2018. Web. https://gdr.openei.org.

Hilbert, M., and 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.

Larson S.: "The hacks that left us exposed in 2017." 2017 Year in Review. CNN Money. 20 Dec. 2017. Web. http://money.cnn.com/2017/12/18/technology/biggest-cyberattacks-of-the-year/index.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.

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.

Seals, T.: "Cyber-attack Volume Doubled in First Half of 2017." Infosecurity Magazine. 11 Aug. 2017. Web. https://www.infosecurity-magazine.com/news/cyberattack-volume-doubled-2017/.

Weers, J. and Anderson, A.: The DOE Geothermal Data Repository and the Future of Geothermal Data, *Proceedings,* 41st Workshop o Geothermal Reservoir Engineering, Stanford University, Stanford, CA (2016).

Weers, J. et al: The Geothermal Data Repository: Five Years of Open Geothermal Data, Benefits to the Community, *GRC Transactions Vol 41*, 41st Geothermal Resources Council Annual Meeting, Salk Lake City, UT (2017).