PROCEEDINGS, 45th Workshop on Geothermal Reservoir Engineering
Stanford University, Stanford, California, February 10-12, 2020
SGP-TR-216

The Data Foundry: Secure Collaboration for the Geothermal Industry

Jon Weers [(a)], and Zach Frone [(b)], Jay Huggins [(a)] and Aaron Vimont [(a)]

[(a)] National Renewable Energy Laboratory, 15013 Denver West Parkway, Golden, CO 80401-3305

[(b)] U.S. Department of Energy, 1000 Independence Ave. SW, Washington D.C. 20004

Keywords: geothermal, data, Foundry, OpenEI, information, EGS, FORGE, GDR, DOE, secure, collaboration, private, storage, transfer, dissemination, access, discoverability

ABSTRACT

The Data Foundry provides secure, cloud-based storage and universal access to digital information, enabling the greater geothermal industry to collaborate seamlessly with the Department of Energy (DOE), national labs, universities, and private organizations. Originally developed to support the EGS Collab project, the Data Foundry has been expanded to support FORGE, EDGE, and other DOE-funded projects, some collaborative, some private, by providing each project with a secure space and the ability to fine-tune individual access controls. In response to user feedback, it now also features improved integration with DOE's Geothermal Data Repository (GDR), to provide a clear and convenient pathway from collaboration to publication, and to register collaborative data projects with well-known data registries like Data.gov and the National Geothermal Data System (NGDS). This paper will explore how recent concerns raised by data-centric, proprietary projects have informed development on the Data Foundry and highlight improvements designed to streamline workflows, improve access control, and promote the timely dissemination of information to the geothermal industry.

1. MULTI-INSTITUTIONAL COLLABORATION

The geothermal industry is trending toward larger, more collaborative projects. Driven in large part by U.S. Department of Energy (DOE) through the funding of projects like EGS Collab and FORGE, the Frontier Observatory for Research in Geothermal Energy, and a desire to offset both cost and risk by encouraging collaboration between national labs, industry, and academia, the geothermal industry is collaborating more frequently on larger projects. Collaborating on large projects, especially those teamed with people from numerous organizations, can be challenging. Institutional policies on user access control and conflicting cyber security practices can introduce barriers to collaboration as individual organizations often place restrictions on external access to their information systems, and a particular tool encouraged for use by one organization may be blocked by another (Weers et al 2018). In order to collaborate effectively, teams must be able to provide each team member with convenient and consistent access to project data and information, regardless of institution.

The volume of data being generated by modern geothermal research and development activities presents its own challenge. Higher resolution instruments, increased sample rates, and lower cost sensors are allowing teams to create more data at higher fidelity. Models and simulations developed by collaborative teams are also increasing in both resolution and complexity (Weers & Anderson 2016). Collaboration around multi-terabyte scale datasets can be challenging as team members must find a way to access these large datasets from their respective institutions. And while tools exist to help move big data resources from one institution to another, collaboration may be limited to those institutions that have super computers, or the special high-throughput infrastructure necessary to receive and process the data.

Barriers to collaboration such as access constraints, conflicting cyber security policies, and institutional bias can limit the pool of potential collaborators, effectively limiting the innovation potential of research and development activities in geothermal. Broad, convenient access to geothermal data was identified by Deloitte in their 2008 Geothermal Risk Mitigation Strategies Report as a means to "reduce the inherent risk in early stages of development and encourage an independent investment market" (Deloitte 2008).

The Data Foundry (https://foundry.openei.org) was developed by the National Renewable Energy Laboratory (NREL) to overcome these challenges, provide consistent and convenient access to data for project partners and reduce barriers to innovation in the geothermal industry by making collaboration easier between national labs, universities, the scientific community, and the geothermal industry (Data Foundry 2020).

2. IDENTIFYING THE NEEDS OF THE COMMUNITY

Originally developed for the EGS Collab project, the Data Foundry was specifically designed to meet the needs of large, collaborative projects. The EGS Collab project, which consists of scientists and engineers from nine US DOE National Laboratories, seven universities, and two private companies, works to establish a collaborative experimental suite of intermediate-scale (~10-20 m) field test beds coupled with stimulation and interwell flow tests. A stepping-stone between laboratory tests and full-scale tests of EGS technology, the EGS Collab project is conducting mid-scale experiments a mile underground in the Sanford Underground Research Facility (SURF), a repurposed gold mine in Lead, SD (Kneafsey et al 2019). The number of collaborators involved in EGS Collab combined with the location of the primary research test bed 1 mile underground present a unique set of challenges in providing collaborative access to project data (Weers and Huggins 2019).

The Data Foundry is the direct result of the specific needs gathered from assessments of the EGS Collab project team and team members from other large scale, collaborative projects. One of the primary needs for a data management tool in the research and development space is the ability to handle any type of data, regardless of format or size. A file management system was selected, as opposed to a relational database or other structured data system, in order to allow users to contribute, collaborate, and access information in whatever format was most convenient for the team at the time. Numerous file management systems are readily available online (e.g. Google Drive) and were considered by the NREL team but were ruled out because they did not meet one or more of the needs of the EGS Collab project (Figure 1) or other data-driven collaborative research and development projects. Most notably, conventional web-based file sharing systems appear to be designed for individual users, allowing personalized customization of shared resources and providing options to support a single-user design model, but creating inconsistency among peers.

Figure 1 Results of a user needs assessment for a collaborative data management system for the EGS Collab project.

2.1 Accessible and Secure

In order to collaborate effectively, data must be accessible to the appropriate teams. Cumbersome login systems or user account management solutions tied to a single organization produce barriers to access and limit the potential for collaboration. All team members should have equal, convenient access to the data they need. That being said, projects occasionally include team members from competing organizations, especially those working with private companies. For these projects, proper data security is paramount. Access should be controlled by both individual and team, allowing key groups of people unfettered access across the project while restricting some project teams to only those data that are needed to accomplish their specific task. All proprietary and business sensitive data should be kept secure, stored in accordance with DOE cyber security guidelines, and managed through appropriate access controls including 2-factor authentication, a means by which the identities of users attempting to access the data are verified twice before granting access.

2.2 Discoverable

Data within the data management system need to be easily discoverable. Even the most critical information will not be used if it cannot be found. Data designated for public dissemination should be made visible to internet search engines through search engine optimization (SEO) and should be described with adequate metadata to enable users to discover it.

2.3 Data Integrity

Projects generating substantial amounts of data can create their own data challenges. If not properly organized, this can negatively impact the discoverability and usability of the data collected. An overabundance of questionable data can increase the difficulty of discovering critical data and dilute the pool of information available to the team. Data in the project space should be of high quality, preferably curated, and clearly versioned to allow future updates without confusion. This requires the adoption of a team-wide, clearly defined data provenance strategy to determine the proper method of versioning existing project data.

2.4 Best Practices and Standards

In order to facilitate collaboration, project members must first speak the same language. Use of standard metadata, data formatting best practices, and previously agreed upon file formats increase the usability of data and reduce the amount of effort needed for data translation and reformatting across the team.

2.5 Outreach and Communication

The success of a project "should be measured not when a project is completed or an experiment concluded, but when scientific and technical information is disseminated" (DOE 2011). A clear path to data publication and dissemination should be defined and allow users of the tool to seamlessly publish data to the GDR, attributing the metadata necessary to promote discovery and understanding of the data.

2.6 Team-Centric Design

Lastly, any collaboration tool must be designed for use by a diverse project team. Each team member should have consistent, universal access to the data they need to complete their work on the project. This includes limiting individual customizations to data organization and views into the data to ensure homogenous representation of the data across all project team members. Sub-teams within the project should have configurable access levels as a team, reducing the need to manage project participants individually.

3. DESIGNED FOR COLLABORATION

The Data Foundry adheres to the design principals outlined by the user needs assessment above and differs from conventional file management tools in that it has been designed to optimize collaboration amongst large, diverse groups. One of the primary differences is the adoption of team-centric design to create a consistent view into data managed by the system. Many of the conventional web-based file sharing systems available today have been designed for individual users, allowing personalized customization when it comes to the organization of shared resources and allowing user access to be prescribed at multiple levels, creating an inconsistent view from user to user (Figure 2).

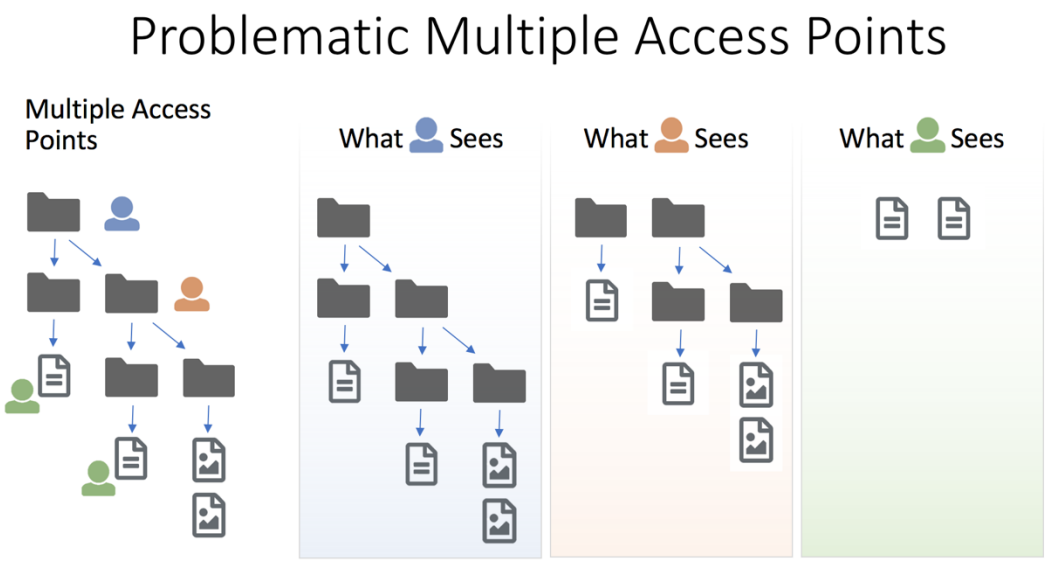

Figure 2 Illustration of a file sharing system with multiple user access points shared by three users (Blue, Red, and Green) and representations of what each user sees.

In the example above, a web-based file sharing system that allows for individual access control and multiple levels has been visualized to show each individual user's view into the file system. Three users, "Blue", "Red", and "Green" each have access to different levels of the file tree, and therefore see different organizational structures for the same information. The left-most column represents the shared file system as it actually exists on the drive, organized by the Blue user. As the organizing user, what Blue sees is an accurate representation of the file system. Red, having access to the tree one level below Blue, sees a slightly different structure. And Green, having direct access to only two files, sees a partial view of the project data in a completely different layout. With each user seeing a different organization of the same information, it can be difficult to collaborate on specific files. Complicating matters more, a reorganization of these assets by one of the users can actually cause data to disappear from the view of another user (Figure 3).

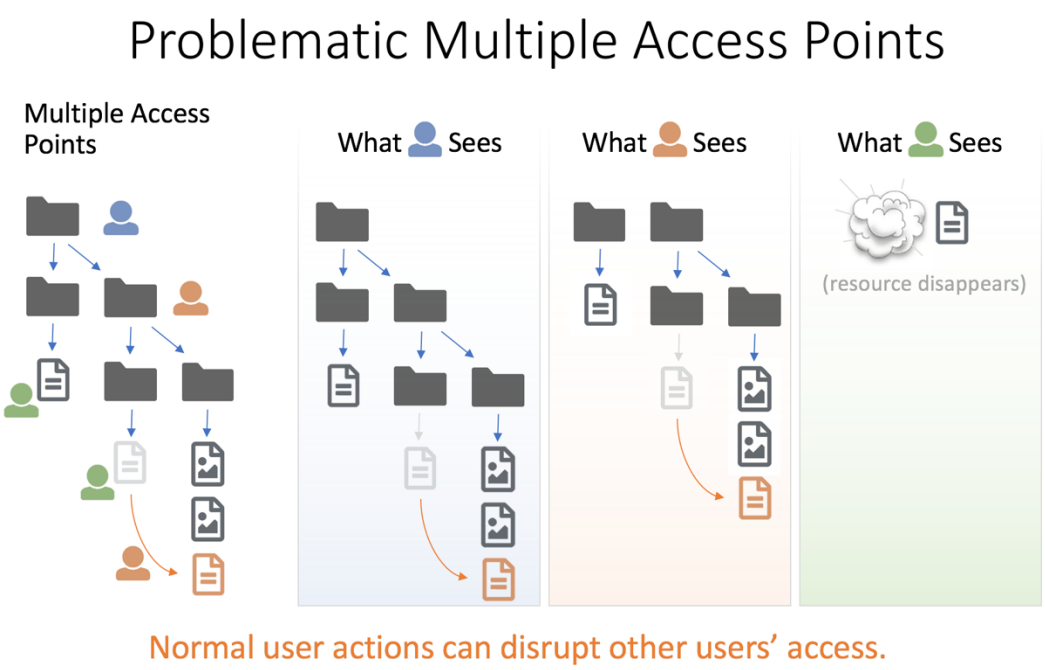

Figure 3 Illustration of a file sharing system with multiple user access points and what each user sees when one user moves a file

In the thought experiment above, the Red user has moved one of the files to a different folder. The Blue user, having access to the tree above, also sees this file move. However, because the file has inadvertently been moved to a folder the Green user does not have access to, the file effectively disappears from the Green user's view. This is a common occurrence in file sharing systems adopting this access paradigm and is a frustration experienced by the users of those systems, often resulting in lost time or data and frequent requests to reshare specific files.

3.1 Consistency from Single Access Points

To avoid these pitfalls, the Data Foundry has adopted a single access point paradigm, effectively limiting access to a project folder at the top level. By limiting access to the top-level folder, the Data Foundry can ensure that each user has equal, universal access to the data resources shared with them and a view into the organizational structure of the data consistent with that of their collaborators (Figure 4). This makes it much easier for teams of people to collaborate on large, complex data structures, and enables them to work together to organize the data into more meaningful and useful structures.

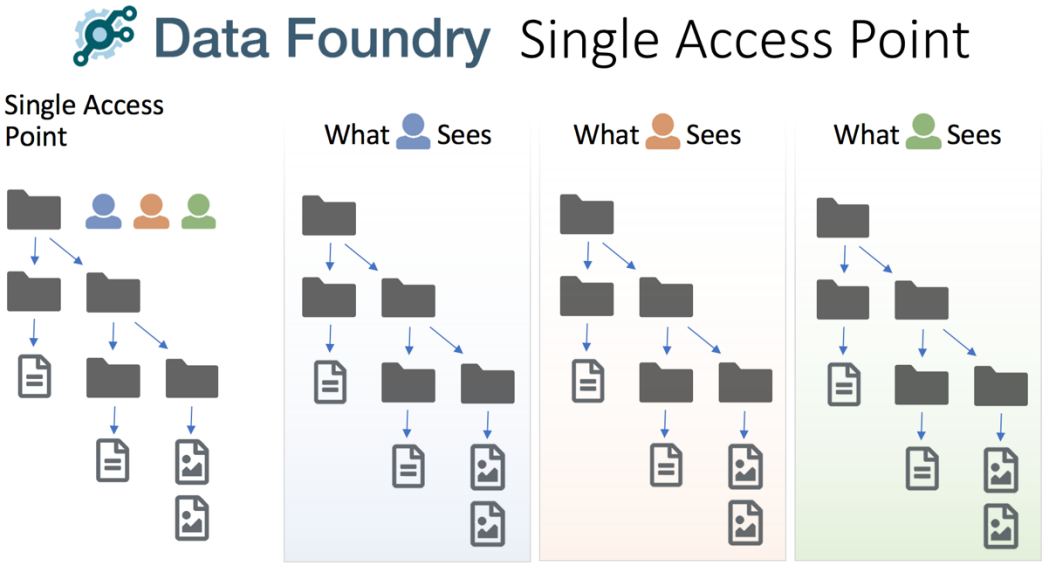

Figure 4 Illustration of the Data Foundry's file sharing system with a single access point providing users with a consistent view

3.2 Configurable Access Management for Teams

Access to the Data Foundry is controlled by the OpenEI authentication system, which allows users to login through an OpenEI or Google account. It does not rely on a single institution's user management system, providing easier collaboration across labs, companies, and other organizations. Users are also assigned one or more spaces. Each space is completely separate from the others, so users from one space cannot see projects from another. On the Data Foundry home page, users can optionally switch between their assigned spaces to only see projects in a single space, simplifying navigation and allowing collaborators to focus on one project at a time.

The Data Foundry has four access levels that can be controlled on a per project basis: View, Contribute, Collaborate, and Manage. Each access level gives a user additional functionality in a project (Figure 5). "View" allows users to view and download resources. "Contribute" allows users to upload new files, in addition to all of the access granted in "View". "Collaborate" builds upon this by also allowing users to move, rename, and delete resources. Finally, "Manage" allows users to assign other users to teams and add those teams to a project in addition to all of the aforementioned actions. User access can be altered at the individual level to give additional permissions to a single user within a team (Figure 6).

Access Levels	View	Contribute	Collaborate	Manage
View projects and resources	✓	✓	✓	✓
Download resources	✓	✓	✓	✓
Upload new files		✓	✓	✓
Move files or folders			✓	✓
Rename resources			✓	✓
Delete resources			✓	✓
Manage user access				✓

Figure 5 Table showing actions allowed at each user access level

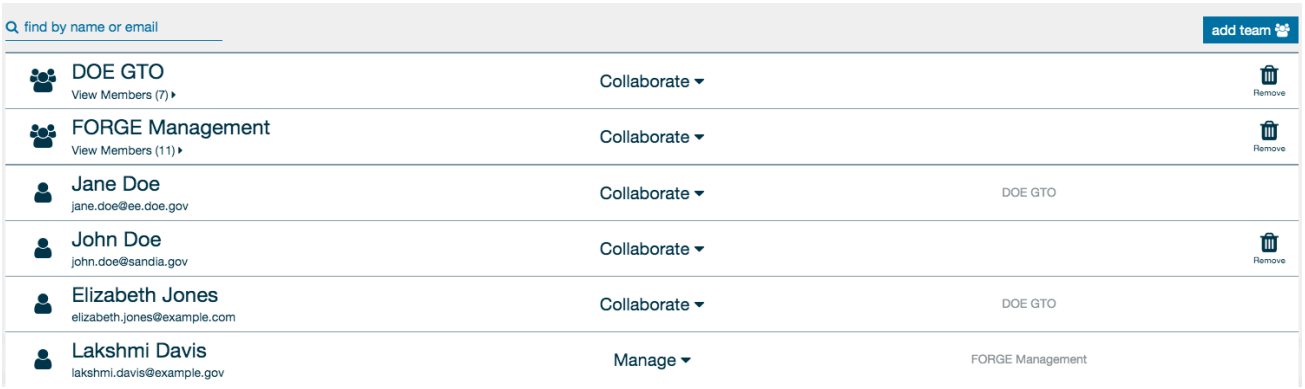

Figure 6 Table listing teams and users and their various access levels in a project. In this example, John Doe has been individually added to the project alongside the "DOE GTO" and "FORGE Management" teams.

3.3 Automated Data Ingestion

Research teams potentially generate large volumes of data around the clock, and they may utilize systems and networks that are robust and reliable, or they may use custom hardware and temporary networks in remote locations, as the EGS Collab team does in the SURF. For either scenario, these teams need unrestricted shared access to the most current data possible, and the Data Foundry currently makes this possible by supporting automated data ingestion via a number of avenues, with safeguards to prevent data loss or corruption, and is extensible enough to easily add support for other protocols and platforms.

Responding directly to client needs, the Data Foundry adopted support for transfers directly from Amazon Web Services Simple Storage Service (AWS S3), Google Drive, and a Globus "endpoint", as well as direct manual uploads.

The peer-to-peer file transfer mechanisms such as Globus are often touted for supporting fast, reliable data transfers. While the Data Foundry does support peer-to-peer file transfer tools, they often require careful and timely configuration on both ends making them difficult to coordinate and impractical at scale for large collaborative projects like EGS Collab. In practice, these types of tools have been difficult to utilize effectively.

Instead, the Data Foundry team has created pipelines to automatically and periodically ingest data from existing staging areas in Amazon S3 and Google Drive. These cloud services feature tremendous upload speeds and virtually unlimited space, but the Data Foundry ingestion system also accommodates slow, unreliable, or intermittent connections by running integrity checks to identify files that may have been inadvertently overwritten, and stores backups of all changed files.

While Data Foundry supports custom scripts in Python for clients to use to perform these uploads, the current recommendation is to use the *rclone* client, which can be easily installed on Mac, PC or Linux, and configured to periodically upload, copy or sync files to numerous cloud services including Amazon S3, Google Drive, Box, Dropbox, and SFTP sites. The Data Foundry can easily be adapted to work with any of these cloud storage systems.

4. STREAMLINED PUBLICATION AND DISSEMINATION OF PUBLIC INFORMATION

The Data Foundry has been integrated with the GDR to make the publication and dissemination of select data to the general public even easier. The GDR and the Data Foundry share a common foundation in the OpenEI cloud architecture, allowing users to access both systems with a single login and enabling data to be passed securely from one system to another. In order to minimize the level of effort needed to craft a GDR data submission from assets on the Data Foundry and reduce the burden of data submission on researchers, integration points between the two platforms were overhauled in 2019. Initiated by clicking the "Send to GDR" link that appears for selected resources (Figure 7), the new, more robust system guides users through the processing of publishing selected resources by attaching them to either a new or existing GDR data submission (Figure 8).

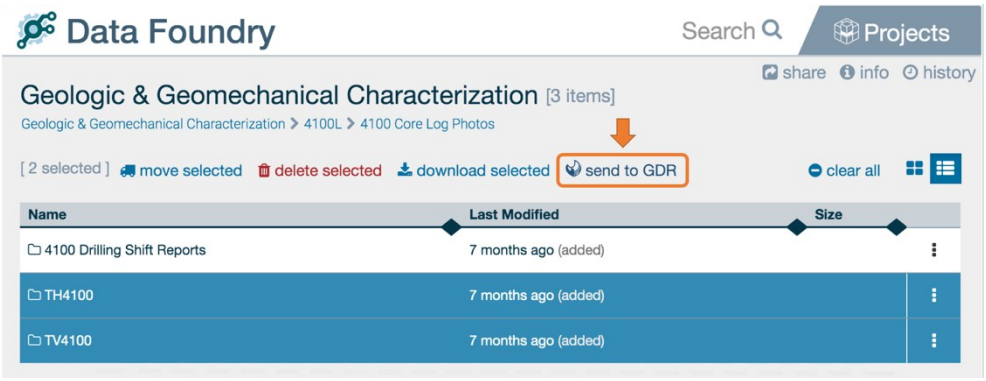

Figure 7 Screenshot showing the "Send to GDR" button (Data Foundry 2020)

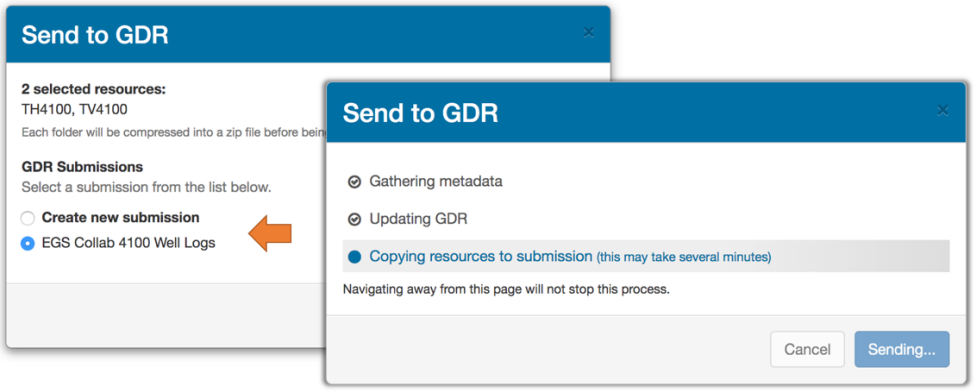

Figure 8 Depiction of "Send to GDR" functionality (Data Foundry 2020)

Metadata sourced from the selected resources and their corresponding project and "space" within the Data Foundry are transmitted to the GDR along with the selected resources themselves to automatically populate the resulting GDR submission. A link to this submission is then provided to the user allowing them to "hop over" to the GDR at their convenience to add any additional supplemental information and complete the submission process, and a note is added to the history of the item in the Data Foundry denoting the date and time of submission to the GDR as well as the name of the person who initiated the submission. This new workflow allows users to collaborate on the formation of a submission and contribute resources from multiple projects within the Data Foundry to one or more submissions on the GDR, providing them with a flexible framework to better organize their data for public dissemination.

4.1 Information Dissemination

Data sent to the GDR for publication through the "Send to GDR' function on the Data Foundry undergoes the exact same review, curation submission, and public dissemination processes as a regular GDR data submission (Weers and Anderson 2016). Once curated and made

publicly available, data published to the GDR from projects on the Data Foundry will be disseminated to the GDR's extensive network of data sharing partners, including Data.gov, the Office of Science and Technical Information's (OSTI) DOE Data Explorer, and Thompson Reuters' Data Citation Index (Weers et al, 2019) (Figure 9).

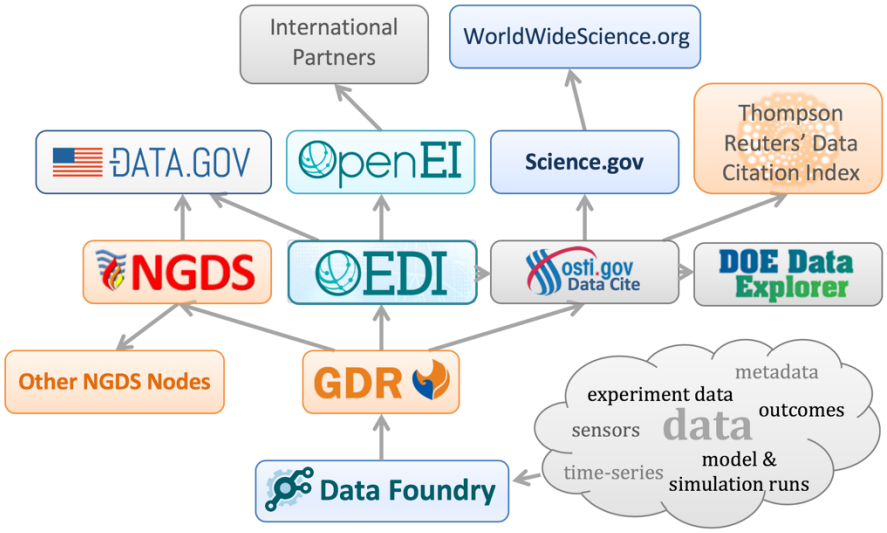

Figure 9: Propagation of Data Foundry data through the network of GDR data partners once published to the GDR.

5. FUTURE IMPROVEMENTS

The NREL team responsible for the Data Foundry continually engages users of the Data Foundry and project stakeholders to solicit improvements to the platform with the goal of further streamlining the data management process to allow users to spend more time on their research and analysis activities. Feedback from the community is incorporated into future development tasks and helps to shape the functionality of the Data Foundry, ensuring that it continues to meet the needs of the geothermal scientific community as they evolve. Future development activities include integrated support for file versioning, API-driven uploads using a token-based authentication system, additional access levels, better tools for collaboration on documents, and the ability to annotate or comment on specific resources.

6. CONCLUSION

The Data Foundry is the direct result of specific data management needs voiced by large, collaborative projects in the DOE Geothermal Technology Office (GTO) portfolio. The team-centric design principles adopted during the development of the Data Foundry reduce barriers to collaboration and ensure that each team member has the same consistent view into data shared by those with similar access. Improved integration with the GDR has provided users with a convenient means of organizing and publishing their collaborative data products and disseminating critical information to the geothermal industrial and scientific communities. The Data Foundry is a critical companion to the GDR, and a vital component in its mission to support geothermal research and development activities, promote collaboration among industry partners, and accelerate the rate of innovation in geothermal technologies.

ACKNOWLEDGEMENTS

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 Data Foundry project, in conjunction with the EGS Collab and Geothermal Data Repository (GDR) projects. The United States Government retains, and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. Some of the research supporting this work took place at the Sanford Underground Research Facility (SURF) in Lead, South Dakota. The assistance of the SURF and its personnel in providing physical access and general logistical and technical support is acknowledged.

REFERENCES

Data Foundry: "Data Foundry: Secure Data Management and Collaboration System." OpenEI: Open Energy Information. National Renewable Energy Laboratory, 30 Jan. 2020. Web. https://foundry.openei.org.

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

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

Dobson, P., Kneafsey, T.J., Blankenship, D., Valladao, C., Morris, J., Knox, H., Schwering, P., White, M., Doe, T., Roggenthen, W., Mattson, E., Podgorney, R., Johnson, T., Ajo-Franklin, J., and E.C. Team (2017): 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.

Heise, J. (2015), The Sanford Underground Research Facility at Homestak, *Journal of Physics: Conference Series*, *606*(1), 26

Kneafsey, T.J., Blankenship, D., Hunter, K., Johnson, T., Ajo-Franklin, J., Schwering, P., Dobson, P., Morris, J., White, Fu, P., Podgorney, R., Roggenthen, W., Doe, T., Mattson, E., Ghassemi, A., Valladao, C., and E. C. Team(2019): EGS Collab Project: Status and Progress, *Proceedings, 44th Workshop o Geothermal Reservoir Engineering*, Stanford University, Stanford, CA (2019).

Knox, H., P. Fu, J. Morris, Y. Guglielmi, V. Vermeul, J. Ajo-Franklin, C. Strickland, T. Johnson, P. Cook, C. Herrick, M. Lee, and E.C. Team (2017), Fracture and Flow Designs for the Collab/Sigma-V Project in *GRC Transactions, Vol. 41,* 41st Geothermal Resources Council Annual Meeting, Salk Lake City, UT (2017).

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

Weers, J., Johnston, B., and Huggins, J.: The EGS Data Collaboration Platform: Enabling Scientific Discovery, *Proceedings, 43rd Workshop on Geothermal Reservoir Engineering*, Stanford University, Stanford, CA (2018).

Weers, J., Anderson, A., and Taverna, N.: 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).