Sustaining the National Geothermal Data System: Considerations for a System Wide Approach and Node Maintenance

M. Lee Allison[1], Cathy Chickering[2], Kim Patten[1], Arlene Anderson[3], and Stephen M. Richard[1]

[1]Arizona Geological Survey, Tucson, AZ
[2]Southern Methodist University, Dallas, TX
[3]U. S. Department of Energy, Geothermal Technologies Office, Washington, DC

Keywords

National Geothermal Data System, NGDS, sustainability, business models

ABSTRACT

Since the 2009 American Recovery and Reinvestment Act the U.S. Department of Energy's Geothermal Technologies Office has funded $33.7 million for multiple data digitization and aggregation projects focused on making vast amounts of geothermal relevant data available to industry for advancing geothermal exploration. These projects are collectively part of the National Geothermal Data System (NGDS), a distributed, networked system for maintaining, sharing, and accessing data in an effort to lower the levelized cost of electricity (LCOE). Determining "who owns" and "who maintains" the NGDS and its data nodes (repositories in the distributed system) is yet to be determined. However, the investment in building and populating the NGDS has been substantial, both in terms of dollars and time; it is critical that this investment be protected by ensuring sustainability of the data, the software and systems, and the accessibility of the data. Only then, will the benefits be fully realized.

To keep this operational system sustainable will require four core elements: continued serving of data and applications; maintenance of system operations; a governance structure; and an effective business model. Each of these presents a number of challenges.

Data being added to the NGDS are not strictly geothermal but data considered relevant to geothermal exploration and development, including vast amounts of oil and gas and groundwater wells, among other data. These are relevant to a broader base of users. By diversifying the client base to other users and other fields, the cost of maintaining core infrastructure can be spread across an array of stakeholders and clients. It is presumed that NGDS will continue to provide free and open access to its data resources.

The next-phase NGDS operation should be structured to eventually pursue revenue streams to help off-set sustainability expenses as necessary and appropriate, potentially including income from: grants and contracts (agencies, foundations, private sector), membership, fees for services (consulting, training, customization, 'app' development), repository services (data, services, apps, models, documents, multimedia), advertisements, fees for premier services or applications, subscriptions to value added services, licenses, contributions and donations, endowments, and sponsorships.

Introduction

The National Geothermal Data System (NGDS) is a collection of geothermal relevant information residing at numerous institutions throughout the country, connected via the internet. Information is being contributed to the NGDS by the state geological surveys, by dozens of public and private research institutions, and by the recipients of federally funded geothermal demonstration projects (Figure 1). Different servers where information resides are referred to as 'nodes' and can be conceptualized as data repositories. The distributed nature of the system enables data to be managed by the institutions and repositories that create the

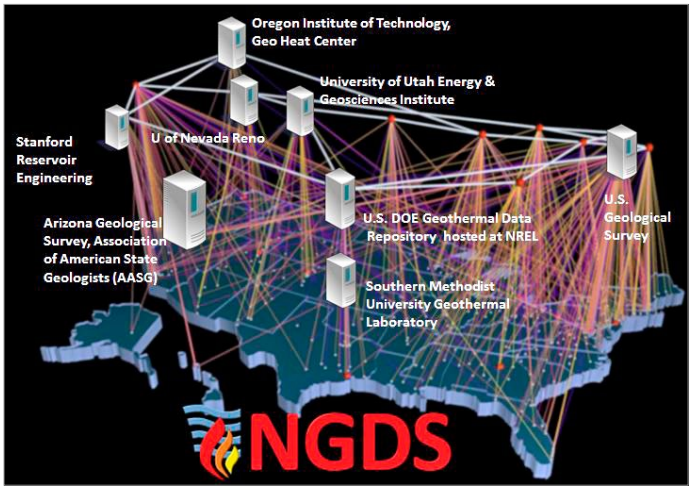

Figure 1. Displays the distributed nature of the NGDS and primary data contributors and nodes.

data, rather than relying on a central 'data czar' where data must be submitted when updated.

The information resources available on these nodes will be 'discoverable' through the use of open source standards and protocols. Specifically the NGDS draws upon the Open Geospatial Consortium's Catalog Services for the Web (CSW) for maintaining records. The CSW is a protocol roughly analogous to a card catalog system detailing the different files available for access with descriptions of who provided the information, where it resides, what type of information it contains, etc. This 'data about the data' is called metadata, and International Organization for Standardization (ISO) standards exist for what minimal information must accompany any file (and geospatial file) in order for it to be discoverable using CSW. One of the advantages of the CSW profile is that it can harvest other CSWs; that is, if two institutions deploy separate CSWs, the other institution's CSW can be 'harvested' thus permitting access to and sharing of those information resources between institutions.

Table 1. Additional information on the Input and Output requirements of each tier.

Tier	Description & Example	Input Requirements	Output Requirements
1	Unstructured Data, for example scanned documents, such as a PDF scan of a scientific report with tables of heat flow data in the appendix.	Upload file to a web accessible location; Create metadata for record.	Accessible via HTTP GET; Metadata points to the URL accessible via CSW
2	Structured (tabular) data not conforming to an NGDS interchange format. Examples include spread sheets, tables, database files etc. such as a Microsoft Excel spreadsheet containing the heat flow data in a format defined by the researcher collecting the data.	Upload files to a web accessible location; Create metadata for the record. Metadata should include descriptive information about the column headers and data types that the documents contain.	Accessible via HTTP GET; Metadata points to the URL accessible via CSW. Optionally, the data may also be exposed as a Web Map Service (WMS) and/or Web Feature Service)WFS).
3	Structured data conforming to an NGDS content model, such as a Comma Separated Value file or MS Excel file that can be transformed for machine consumption in a specified format.	Upload files conforming to standard format to a web accessible location; Create metadata pointing at all endpoints (because the content model defines entity/attribute information the creation of metadata is less cumbersome).	Accessible via HTTP GET; Data exposed to the internet using Web Feature Services/Web Map Services; Metadata points to all endpoints accessible via CSW.

The data within the NGDS include 'unstructured data', such as scanned PDFs of scientific papers or reports, notes or images etc. While not ideal for sharing information, compared to no information, unstructured data is certainly 'good' to have. Additionally, and perhaps most importantly, the NGDS also includes 'structured data', such as spreadsheets, database files or GIS files; this structured data is generally 'better', meaning more readily usable, than unstructured data. Finally, the NGDS promotes the interoperability of data between data providers, through the deployment of 'interchange models' for common data types, e.g. well logs, borehole temperature data, aqueous chemistry, etc. This level of interoperability has been designed to reduce the Pareto Principle (80/20 rule) when applied to scientific procedures. Anecdotally, it is estimated that the current scientist/researcher, spends approximately 80% of their time researching and manipulating data and only 20% of their time analyzing data. The vision of the NGDS is to reverse this for geothermal exploration allowing industry to invest the majority of their time developing the geothermal resource.

These different types of information contributions to the NGDS have been referred to as Tier 1 through 3 data:

- Tier One – unstructured data, with metadata describing the location and type of information contained in the file,
- Tier Two – structured data, with metadata describing the location and type of information contained in the file,
- Tier Three – structured data in a standardized defined format, with metadata describing the location and type of information contained in the file.

For example, if the temperature-at-depth and heat flow data available from one of the universities, such as the Southern Methodist University (SMU) Node of the NGDS, is in the identical format as the temperature-at-depth and heat flow data provided by different state geological surveys, software can be written to poll all of the NGDS nodes and combine it into a single view. Applications such as Geothermal Prospector, a project of the National Renewable Energy Lab (NREL), aim to do just this, using the OGC protocol called Web Feature Services (WFS). In order for WFS to work as intended, the information providers have to present their information in a predefined layout so that the software will understand where different fields of information (latitude, longitude, temperature, depth, units of measure or spatial referencing system used, etc.) reside within each file.

The task of putting information into a predefined standardized layout can vary from simple to highly complex. Currently, most data providers to the NGDS can contribute any Tier of data, although, Tier 3 data are most valuable (if applicable) as they can then interact with other data sets in the system. Tutorials on data interoperability and standardizing data formats are available from the NGDS web site at www.GeothermalData.org.

Component Projects

According to the Department of Energy, Energy Efficiency, and Renewable Energy, Geothermal Technologies Office (GTO), there are five main projects within the NGDS, three funded cooperative agreements and two interagency agreements. These include:

1. National Geothermal Data System Architecture, Design, Testing, and Maintenance, PI Institution: Boise State University (BSU), Synopsis: responsible for the systems architecture and data compilation from the University of Nevada Reno, Oregon Institute of Technology, University of Utah, and Stanford University.
2. Association of American State Geologists, State Geological Survey Contributions to the NGDS, PI Institution: Arizona Geological Survey (AZGS), Synopsis: responsible for the data digitization and integration from the fifty state geological surveys in addition to a small amount of new data collection.
3. Heat Flow Database Expansion for NGDS Data Development, Collection, and Maintenance, PI Institution: Southern

Methodist University (SMU), Synopsis: responsible for updating the SMU Heat Flow database with contributions from Siemens Corporate Technology (SCT), Bureau of Economic Geology at the University of Texas at Austin, Cornell Energy Institute at Cornell University, Geothermal Resources Council, MLKay Technologies, Texas Tech University, and University of North Dakota. All institutions are providing data with the exception of SCT which is providing database and data analysis work.

4. Geothermal Resource Classification, PI Institution: USGS, Synopsis: an update to the 2008 Geothermal Resource Classification.
5. Geothermal Data Repository, PI Institution: National Renewable Energy Laboratory, Synopsis: responsible for curating and hosting the data from GTO funded projects for submission to the NGDS.

Why NGDS?

In March of 2011, DOE convened a panel of geothermal experts in Albuquerque, NM for a guided discussion on the future of geothermal energy in the U.S (U.S. DOE, Office of Energy Efficiency and Renewable Energy, 2011). One of the recommendations from the 15 panelists was that DOE focus its resources on reducing the cost of confirming known hydrothermal resources and identifying undiscovered hydrothermal resources. In order to do this, the panel suggested developing an inventory of prospects using existing technology. Also in 2011, the International Energy Agency released the *Technology Roadmap: Geothermal Heat and Power* which called for "develop[ing] publicly available databases, protocols, and tools for geothermal resource assessment and ongoing reservoir management to help spread expertise and accelerate development" (International Energy Agency, 2011). When completed, NGDS will help achieve both reports' recommendations.

Finally, in an effort to ensure that federally funded research is "available to and useful for the public, industry, and scientific community," the Executive Office of the President, Office of Science and Technology Policy, released a Memorandum for the Heads of Executive Departments and Agencies *Increasing Access to the Results of Federally Funded Scientific Research* in February 2013, that requires agencies with research budgets over $100 million to ensure that "…the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community. Such results include peer-reviewed publications and digital data" (Holdren, 2013). The memo further requires that agencies ensure that publications and metadata produced with federal funds are stored in an archival solution that:

1. Provides for long-term preservation and access to the content without charge
2. Uses standards, widely available and, to the extent possible, nonproprietary archival formats for text and associated content (e.g. images, video, supporting data)

Thus, the NGDS vision of shared and easy data access can assist DOE in completing these requirements.

Why Sustainability of the NGDS Matters

One of largest barriers to entry in geothermal projects is the upfront capital outlay and the inherent risks associated with the potential for drilling a dry or under-producing well. In February of 2013, *Scientific American* quoted Susan Petty, President and Chief Technology officer of AltaRock Energy as saying that "The risk involved in geothermal prospecting sets the industry apart from other renewables" (Ferguson, 2013). Clearly, knowing *where to drill* can reduce the risk and thus increase the market penetration of commercial scale geothermal power.

The Department of Energy, primarily through funds allocated under the American Recovery and Reinvestment Act of 2009, has invested $33.7 million dollars in developing the infrastructure for the NGDS and in populating it with high quality information needed to reduce the risk associated with geothermal development (U.S. Department of Energy, Geothermal Technologies Office, n.d.). This is separate from the funding that was directed to geothermal demonstration projects, which included a mandate for contribution of data to the NGDS under the terms of those demonstration project grant contracts.

Advancements in drilling techniques, in geothermal reservoir engineering, and in temperature data made available since the boom in U.S. natural gas drilling are just some of the information that is new within the last few years. The over $300 million in federally funded geothermal demonstration projects are expected to generate significant new understanding that could be leveraged in this developing industry (U.S. Department of Energy, Geothermal Technologies Office, n.d.). The Geothermal Data Repository (GDR) is striving to become a node on the NGDS that will house the data collected from these projects. In addition to the 'new' information, the pioneering researchers in geothermal energy are approaching retirement; a thoughtful approach to making legacy data available will ensure that their research will be available for further analysis and study in the years to come.

If the justification for this significant investment in the NGDS is the importance of making key information about subsurface resources available to potential project developers, ensuring the information can be updated and will be accessible to users for years to come is equally vital.

The NGDS vision is such that the functions, accessibility, and availability of data will encourage continued participation within the NGDS, from all end users – data consumers, data providers, and application developers. Once the protocols and standards are in place in each data node, additional data added to each database are automatically promulgated to the system. As each node adds to is data repositories, the NGDS functions become increasingly valuable to it. Each data provider will have created a value added service that is transportable and scalable to cover all data in its possession. Thus, there are benefits to each participant to continue to add data to the system and maintain it. In the long term, the data network is expected to reach a 'tipping point' at which it becomes like a data equivalent to the World Wide Web – where stakeholders will maintain the function because it is expected and it fulfills critical needs. Applying this vision to the NGDS, it also opens the door for additional data providers external to geothermal development, thus wholly increasing the value of NGDS and its

underlying data integration platform USGIN (further discussed in the System Operations section).

Factors Affecting Sustainability of Information Resources

Just as a chain is only as strong as its weakest link, there are numerous opportunities for the process to 'break'. For the NGDS to be sustainable for years to come, resources must be directed at:

1. Keeping the information current and meaningful ('Content Sustainability'),
2. Governance, or organizing and managing the system (or system of systems) responsible for the content and system operation of the NGDS in a cohesive and effective manner ('Organization and Management'),
3. Ensuring the systems where the data resides can be accessed, operated and maintained for years to come ('System Operation Sustainability'), and
4. Establishing a business strategy, including a funding mechanism, for ongoing operation ('Business Models and Plans').

A strong sustainability plan will encompass all four of these challenges.

1. Content Sustainability Considerations

The data contributed to the NGDS must be kept current, if users are to find it valuable for years to come; this is, in fact the strength of the distributed network. The nodes of the NGDS must continue to be populated by current sources of raw data: public and private industry, governmental agencies, and educational institutions. Measures that will help ensure NGDS content stays current include:

1. Continue the requirement that all recipients of GTO funding must contribute certain types of data to the Geothermal Data Repository as a contingency of their federal support.
2. Require or financially incent the next incarnation of data contribution to adhere to guidelines for a specified type of data contribution (Tier 2 and Tier 3, versus Tier 1 as described above)
3. Expand this model to other sources of public funding, including the National Science Foundation (NSF), and other program areas of DOE such as the fossil fuel driven initiatives, USDA Rural Development grant recipients, etc. Additionally, research and development that falls within Dr. Holdren's Memo to heads of scientific research agencies may be encouraged to contribute data.

New data contributions by private industry not supported through public funds are less likely to be forthcoming, without either regulatory mandate or significant financial incentive. When a private organization has key information not readily available elsewhere within NGDS, compensation for NGDS contribution should be considered. The compensation would not necessarily need to be in a direct outlay of funds, but could include tax incentives and/or preferential access to content or services on the NGDS in exchange for voluntary participation.

In addition to populating the NGDS with completely new contributions, there should be a means for previous contributors to continue to add to their data collections. As it becomes possible to aggregate related information coming from different sources, new analysis and/or new interpretations of existing data are expected. Incorporating these additions to existing data will be crucial to maximizing the benefit of the investment in the NGDS. As part of this, nodes on the NGDS must enable the ability to determine when updates were last made and the format of the previous submissions. General users of the system (as well as data contributors) should have the ability to identify the source of the information as well as the date and information regarding the collection or analysis.

In the case of the SMU Node of the NGDS, the data contributors will have the ability to upload files in the same format as used previously to update or augment previous submissions. They will have specially designated user IDs that will enable them to see the last successfully uploaded file and/or upload new files. Once satisfied that the data is validated, they can elect to 'publish' it, thereby making it available to the rest of the NGDS and the public at large. Contributors to the SMU Node upload their files as 'Tier 2' structured data, which is then transformed by custom software written by Siemens Corporate Technology into one or more of the 'Tier 3' layouts specified for use in exchanging data via WFS on the NGDS. Other nodes of the NGDS will handle this conversion differently, some of them through the use of installation of a Department of Energy funded software product referred to as 'Node-in-a-Box'. Developed on CKAN 2.0, this open source product will aid data contributors in translating data from a native layout into one or more of the formats specified for transmitting Tier 3 data via the NGDS.

Even with the use of translation software like the CKAN based Node-in-a-Box, some information may not be able to be readily reformatted into 'Tier 3'data layout without significant work effort. For this reason, future recipients of federal funds that will be contributing data to the NGDS should be notified of the specific layouts expected at the time of contract negotiation, to minimize this difficulty.

2. Organization and Management

As mentioned previously, there is no 'one' governance system that comprises the NGDS, nor is there one centralized responsible party. Rather, it is a loose collection of systems and organizations, each with their own methods of operation and management. USGIN is a framework of adopted standards, protocols, and practices, which weaves the NGDS components together into a federated distributed network. NGDS is currently the largest 'client' of USGIN.

3. System Operations Sustainability

NGDS is designed as an open-source, web-based, nationally distributed (vs. centralized) system with diverse users and providers, making it scalable and transferable to data providers and users in other fields.

NGDS is built upon the data-sharing platform, U.S. Geoscience Information Network (USGIN) which is series of open-source standards, protocols, and specifications funded by the National Science Foundation from 2008-2012. It is a collaboration between the U.S. Geological Survey and Association of American State

Geologists resulting from a 2007 workshop recommending the creation of a distributed, national "Geological Information Network," of digital Earth Science data using common standards and protocols, preserving ownership, credit, and control of data, and building on existing data systems (Allison & Dickinson, 2008). USGIN enables users to efficiently find, access, and share geoscience data by registering (and discovering) data using standardized catalog services and metadata. Because USGIN is a network, it is largely intangible and is rather defined by the protocols, interchange formats, and conventions that enable its operation (Richard, 2013). Documents developed under USGIN are now under review and maintained in an online repository accessible to the public (https://github.com/usgin). Using these documents, anyone can replicate, maintain, and update the data integration framework for NGDS. The data system approach of individual web hosts and repositories, in conjunction with the networking capabilities that have developed and been implemented in NGDS are expected to be used subsequently for more and broader data than geothermal.

Once the protocols and standards are in place in each data node, additional data added to each data base are automatically promulgated to the system. As each state adds to its data bases, the NGDS functions become increasingly valuable to it. Each data provider will have created a value-added service that is transportable and scalable to cover all data in its possession. Thus, there are benefits to each participant to continue to add data to the system and maintain it.

The systems where the data resides must be maintainable. The use of the term 'systems' encompasses the physical hardware, the operating system software and network connectivity. This is complicated by the fact that there is no 'one' system or even one organization responsible for the ongoing maintenance of the various systems comprising the NGDS. AZGS is charged with the maintenance of the AASG server(s); DOE and NREL will maintain the Geothermal Data Repository; SMU will maintain their node (Geothermal Data Aggregation, GTDA), etc. While this may initially appear as a fault of the system, this is also the strength of the system, because it eliminates a single point of failure.

Making use of open source systems and/or widely used commercial products will help provide interoperability, which is a necessary component for ongoing sustainability in a distributed system environment, but it does not fully address the myriad of issues that must be considered. In addition to selecting open source options where practical, utilizing 'loosely-coupled' components also aids sustainability. This means, that front-end, or client side applications, are not dependent upon an individual back-end, or server side application. It is similar to the notion that a car owner should be able to replace their car battery without needing to replace the entire engine, and without needing to even understand how the engine works. To continue with the car analogy, just as there is an interstate highway system with standard 'rules of the road' understood by all types of drivers in all sorts of vehicles, even when passing through different state and local jurisdictions, the network that communicates among the NGDS nodes must be open to participation from a myriad of locations and platforms, provided they make use of certain protocols when communicating information.

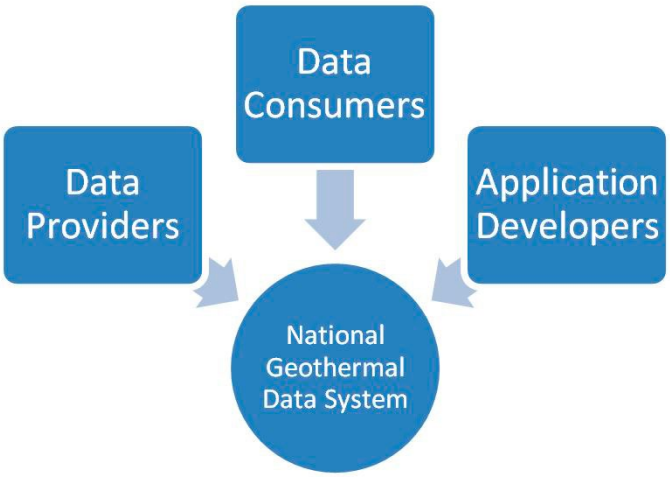

Figure 2. Categories of User Groups.

System Users

In addition to operating system software, hardware and accessibility, user administration is another element of Systems Operations that must be considered with regard to ongoing sustainability. The system is designed to encourage use by not requiring logon information for data retrieval. User communities fall into one (or in some cases more than one) of the following categories, each with their own set of requirements: Data Consumers, Data Providers, and Application Developers (Figure 2).

Data Consumers: Individuals and/or organizations interested in querying the information within the NGDS. This could include anyone from a professor, to a legislator, to a business executive to an interested member of the public. In large part, 'read only' type access to the information on the NGDS should be freely available. Use of this approach will minimize the need for ongoing support within NGDS for establishing and maintaining user credentials. Individual nodes, or external data providers, may choose to implement logon-credentials for accessing their data, but discovery of the data should remain free and accessible.

Data Providers: Individuals and/or organizations with data (Tier 1, 2 or 3) to load onto a node of the NGDS for access by general users OR individuals/organizations who would like to contribute data by becoming a node. Some nodes may have the need to distinguish among roles within this category, such as those who may add to data collections from those who may remove or replace data. In any event, administering this type of user community is likely to require knowledge of the individual(s) contributing data, where they are located, under whose authority they are acting, etc. Maintaining this function will be ongoing and will likely require helpdesk type support capabilities or a 'node administrator.' This could range from resetting passwords, to providing guidance on file uploading procedures, sources of documentation for software questions, etc. The node through which contributors upload information will likely need to be prepared to support this function. Users who wish to host their own node on the NGDS could potentially require some assistance by a centralized group. There may also be the requirement of a long-term 'broker' or individual that can assist with data transformation. Currently this is being handled by the individual

nodes. The vision, though, is that data providers will recognize the value of portraying their data in Tier 3 format for interoperability. This is not an unlikely vision as the World Wide Web has similar standards and protocols for 'participating' in the web, such as use of numerical IP addresses for identifying servers, HTTP protocols for communication between clients and servers as well as authentication of users, registered 'domain names', a defined textual markup language (HTML) etc.

Application Developers: Individuals or organizations interested in developing additional functionality to the NGDS and maintaining its open source code. These users are likely to have more information technology experience than either the Data Consumer or the Data Provider. While it is not necessary to know where these application developers are located or what they will do with the data, it is helpful to have anticipated questions about the data format layouts and communication protocols used. A variety of helpdesk type support might be necessary, but could be reduced by well written easily accessible documentation, or through the development of online discussion forums as are typical in developer environments, such as GitHub repositories.

Ecosystem of Users

NGDS and its underlying architecture, USGIN, have been developed with the conception of an 'ecosystem' of users that evolve with time and development. The success of the ecosystem depends on the longevity of the software and all that relates to it, such as the data formats, protocols, etc. NGDS was specifically built on open-source and commonly used standards and protocols in order facilitate that opportunity for the software to evolve. Parties involved in the ecosystem can improve parts of the system by dedicating resources toward that end as fits their needs, especially if well documented. Crucial factors that enable this are:

- Usability and quality documentation which makes the system usable from the standpoint of the data provider and data consumer. Systems that are less easy to understand or otherwise present a barrier to entry without significant reward are less likely to flourish.
- Code quality and sound system documentation make the system usable from the standpoint of the software/application developer. Systems that are easy to build on are more likely to attract developers who always have the option of adopting other systems or building from scratch.

Often, at the onset, centrally guided efforts are necessary – such has been done with the NGDS – to develop the initial software and protocols. Long-term, however, an organization that aggregates the activities of long-term maintenance (without necessarily taking full responsibility for executing all activities) is very beneficial to the sustainability of the NGDS. As can be seen with other similar software ecosystems (e.g. The Apache Foundation, The Mozilla Foundation) they thrive based on the interest and involvement of actors in the ecosystem, but only by addressing the needs of the ecosystem.

Operational Standards for Major Nodes

Even though each node will likely make their own choices regarding computer manufacture, operating system, network providers, etc., the responsibility for selecting the communication protocols must be coordinated and based on certain standards and protocols in order to provide interoperability. Although there is currently no provision for ongoing federal support of individual nodes on the NGDS, it would be worth considering linking financial support for maintenance of several major NGDS nodes with meeting certain minimum recommended standards. Minimum standards could potentially include some or all of the following:

- Minimum storage capacity and provisions for added growth of disk space
- Minimum network connectivity bandwidth
- Regular backup and maintenance procedures according to a schedule, with off-site storage of backups
- Codified disaster recovery plan with alternate facilities (For those organizations without an alternate backup facility, they could agree to serve as a backup facility with another node in exchange for the same)
- Physical system security precautions
- Software system security precautions
- User administration support (for data contributors to that specific node and/or to block inappropriate use from within their network domain)
- Commitment to participate in the NGDS using the standards and protocols established by the lead team
- Provisions for the case where the node may become inactive so that information is not lost

The potential for implementing a Service Level Agreement (SLA) is a common vehicle in the Information Technology world for specifying such expectations including system quality attributes such as interoperability, reliability, availability, recoverability, performance, integrity, confidentiality, etc.

Example of Node Operations – the SMU Node

SMU Geothermal Laboratory hosts one of the NGDS nodes at their primary campus in Dallas, TX. They developed a written sustainability plan encompassing the following aspects of sustainability:

1. The geothermal data submitted by the 7 institutions providing data to the SMU Node
2. The software developed by Siemens for the project, also called 'GTDA'
3. The servers, which consist of system software and hardware, and
4. The network connectivity

A key aspect of the system is the ability for each of the data providers to upload additional data using the same format as initially submitted to Siemens. The GTDA software will add to or replace previous uploads using these import routines. SMU's Office of Information Technology (OIT), has operated and managed the hardware and system software since the early stages of the project. They will continue to manage the environment for the next 10 years, based on the following expectation:

Item	Function Description	Comments
1	3 Servers (VM) for Application, Web, and Database	Each with 16GB memory & Quad core processors. 2 servers with 500GB disk storage each and 1 server with 3 TB disk storage on shared Storage Array. Located in physically secured facility with temperature and humidity controls.
2	Routine Server Hardware and operating and system software maintenance	Software environment includes Apache Redhat Enterprise Linux (RHEL), PostgreSQL, GeoPortal, GeoServer, and additional open source components. Maintenance inherently limited to those made available by the open source community. Updates to VM hardware environment to be determined by SMU Office of Information Technology (OIT), as needed to maintain reliability and efficiency.
3	Storage Capacity increase allowance of up to 10% per year (simple, not compounded)	For a 10 year project life, up to 100% additional space on the shared array, or an additional 3 TB, is within the agreed upon allowance. Should storage capacity be required in excess of the growth allowance, OIT will allocate additional space at an incremental cost not to exceed $500 per TB.
4	Support of Geothermal Data Aggregation (GTDA) Application developed by Siemens Corporate Technology	In the event that updates or 'patches' to the application are provided by Siemens and/or the NGDS administrators, SMU OIT will install them using the instructions provided by Siemens and/or the NGDS administrators. SMU OIT will assist with loading of large batch updates to records, provided they are in the same file layout/format/structure as currently supported by the applications, using the instructions provided by Siemens. Individual updates, including adding, deleting and modifying existing records will be handled by the users.
5	User Administration Support	SMU Geothermal Laboratory personnel and SMU OIT personnel will both have administration authority to set and modify permission levels for a minimum of 2 user IDs per data provider (BEG at the University of Texas, Cornell University, Geothermal Resources Council, MLKay Technologies, SMU Geothermal Laboratory, Texas Tech University, and University of North Dakota). SMU Geothermal Laboratory and SMU OIT personnel will have the capability to block users, including general public users accessing the system via a federated logon as needed.
6	Internet Network Connectivity	Assumes use of the SMU Internet Connectivity in use at that time and accessibility by the general public via the internet. The assumption is that the speed and capacity of the connection will be at least as good as connectivity in place in 2012. SMU OIT may make changes as needed to maintain reliability, efficiency and to protect internal SMU network and server resources.
7	Data Backup for Disaster Recovery	Assumes max data of 3TB with up to 10% per year increases in capacity. Backups are done via the network to off-site facility. Incremental data backups are performed 4 days per week and weekly backups are conducted each weekend. Backups are retained for 4 weeks. Battery uninterruptible power supply and motor generator backup on premises. Alternate data center available 20 miles from main campus in Plano, Texas. SMU OIT may adjust to the backup schedule to maintain consistency with the backup schedule of other SMU resources and/or compliance with the SMU Computing and Communication Policy.

4. Business Models and Plans

The American Recovery and Reinvestment Act of 2009 was the primary means through which the various elements of the NGDS were created, with oversight provided by the EERE's Geothermal Technology Program. The awards expire as early as September 2013, in the case of the SMU project, with no confirmation of continued federal funding. Various business models and plans used by other nationwide geo-information data and service providers dedicated to other specialties (seismology, meteorology, hydrology, etc.) are under review to determine which methods have been successful in long term funding support for the inherent costs of maintaining a system of the size of the NGDS and its major nodes. The National Science Foundation supports a number of these other efforts, but there are other models worth consideration involving commercialization of some portions of the system.

While it is unlikely that something built primarily by government funding would be entirely privatized, establishment of a non-profit entity would open up possibilities for membership support, advertising income, custom services available for a fee, as well as support from grants and/or contracts.

Data being added to NGDS are not strictly geothermal only but data relevant to geothermal exploration and development, including vast amounts of oil and gas and groundwater wells and other data. These are relevant to a broader base of users. By diversifying the client base of NGDS to other users and other fields, the cost of maintaining core infrastructure can be spread across an array of stakeholders, many of which are larger than the DOE Geothermal program. It is presumed that NGDS will continue to provide free and open access to its data resources.

The next-phase NGDS operation should be structured to eventually pursue revenue streams as necessary from grants and contracts (agencies, foundations, private sector), membership, fees for services (consulting, training, customization, apps), repository services (data, services, apps, models, documents, multimedia), advertisements, fees for premier services or applications, subscriptions to value added services, licenses, contributions and donations, endowments, and sponsorships.

Additionally, several of the institutions contributing data to the NGDS are involved in research that crosses U.S. boundaries and interest has been expressed in how to leverage the U.S. NGDS efforts for global impact. This is another area that could open up possibility for a source of financial support of the NGDS.

As of this writing, the final determination as to who "owns" the NGDS or how it should be managed after the ARRA funding has expired has not been finalized. One concept is to treat NGDS as a DOE-owned asset that can be managed by an external entity under contract or grant from the DOE. NGDS is run on the US Geoscience Information Network data integration framework of standards, protocols, and practices. USGIN is expected to be established as a non-profit entity and could maintain NGDS as one of a number of loosely coupled client applications on behalf of DOE. Alternately, NGDS could operate independently or within another organization such as a professional society or industry trade association, and use USGIN just like other it uses other open source software (e.g. CKAN), or code repositories (e.g. GitHub) and standards (e.g. OGC, ISO).

Case Studies of Successful Sustainability Programs in Data Management

In order to ensure that NGDS has the highest chance of sustainability success, research on a variety of institutions and organizational models is ongoing. We are reviewing multiple funding options as well as organizational structures from non-profit organization to federally funded data centers. Case studies include the governance structure, staffing situation, and funding. Some examples of organizations investigated include, the W3C Consortium, DataONE, Unidata, Apache Software Foundation, Open Knowledge Foundation, Open Geospatial Consortium, Open Topography, and OneGeology.

Example information drawn from a case study of the Apache Software Foundation includes a general background of the foundation, its history, governance structure, funding, staff, and sustainability. ASF demonstrates with relatively small amounts of funding – GuideStar indicates that ASF operated for approximately $405,000 from May 2010 to April 2011 – they were still able to produce and maintain projects that serve enterprise-level solutions. This is primarily through a commitment to collaborative open source software development which has resulted in the delivery of enterprise-grade, freely available software products that attract a large community of users.

Conclusion

NGDS is intended to be an operational data system rather than a research experiment or test bed. It is designed to function as ongoing infrastructure, meeting the needs of the geothermal community as a working tool. Once the initial design/build/populate/deploy phase is complete, to the system must transition into a sustainable resource not subject to the vagaries of federal funding for its maintenance.

NGDS was designed from the start to be used by a wide range of spatial data users, particularly in the geosciences and resource communities, but also land managers and researchers. The strategy of making NGDS scalable and transportable is being proven out by its adoption by a growing cadre ("ecosystem") of users outside the geothermal field, for example, the Western Regional Partnership has agreed to serve elements of its cache of more than 10,000 data layers in USGIN style web assets. Additional interest has been shown by the upstream oil & gas industry through their Energistics Consortium. Thus, with multiple adopters and interested adopters, offering the potential for widespread and diverse financial and technical support, and ensured and expanded data and service providers. Adoption of a governing structure and business model are the next steps in preparing NGDS to take its place as an established resource and community asset.

Acknowledgements and Disclaimers

This material is based upon work supported by the Department of Energy, Office of Energy Efficiency and Renewable Energy, Geothermal Technologies Office, under Award Numbers DE-EE0001120, DE-EE0002850, and DE-EE0002852.

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.

References

Allison, M. L. & Dickinson, T. L., 2008. *A workshop on the role of state geological surveys and U.S. Geological Survey in a geological information system for the nation,* Tucson, AZ: Arizona Geological Survey.

Ferguson, W., 2013. *New Geothermal Data System Could Open Up Clean-Energy Reserves.* [Online] Available at: Permanent Address: http://www.scientificamerican.com/article.cfm?id=new-geothermal-data-system [Accessed 1 May 2013].

GuideStar, 2013, Apache Software Foundation [Online] Available at: http://www.guidestar.org/organizations/47-0825376/apache-software-foundation.aspx [Accessed June 5, 2013].

Holdren, J. P., 2013. *Increasing Access to the Results of Federally Funded Scientific Research,* Washington, D.C.: Executive Office of the President, Office of Science and Technology Policy.

International Energy Agency, 2011. *Technology Roadmap: Geothermal Heat and Power,* Paris, France: International Energy Agency.

Richard, S. M., 2013. *USGIN White Paper - in development,* Tucson: Arizona Geological Survey.

U.S. Department of Energy, Geothermal Technologies Office, n.d. *About the Geothermal Technologies Office.* [Online] Available at: http://www1.eere.energy.gov/geothermal/about.html [Accessed 3 May 2013].

U.S. DOE, Office of Energy Efficiency and Renewable Energy, 2011. *Geothermal Technologies Program Blue Ribbon Panel Recommendations,* Washington, D.C.: U.S. DOE.