Proceedings World Geothermal Congress 2015
Melbourne, Australia, 19-25 April 2015

Open Access to Geoscience Data for Exploration and Assessment

M. Lee Allison[1], Stephen M. Richard[1], Kim Patten[1], Christy Caudill-Daugherty[1], and Arlene Anderson[2]

Mailing address, [1]Arizona Geological Survey, 416 W. Congress St., #100, Tucson, Arizona, USA

[2]U.S. Department of Energy, Geothermal Technologies Office, 1000 Independence Ave., Washington DC, USA

E-mail address, Lee.allison@azgs.az.gov; steve.richard@azgs.az.gov; kim.patten@azgs.az.gov; christy.caudill@azgs.az.gov; Arlene.Anderson@ee.doe.gov

Keywords: Data, open access, open-source, digital exploration

ABSTRACT

The U.S. National Geothermal Data System's (NGDS – www.geothermaldata.org) provides free open access to millions of data records, maps, and reports, sharing relevant geoscience and land use data to propel geothermal development and production in the U.S. Since the NGDS is built using the Geoscience Information Network data integration framework the system is compliant with international standards and protocols, scalable, extensible, and can be deployed throughout the world. NGDS currently serves information from hundreds of the U.S. Department of Energy's sponsored development and research projects and geologic data feeds from 60+ data providers in all 50 states, using free and open source software, in a federated system where data owners maintain control of their data. This interactive online system is opening new exploration opportunities and shortening project development by making data easily discoverable, accessible, and interoperable at no cost to users.

1. INTRODUCTION

Since the 2009 American Recovery and Reinvestment Act (ARRA) 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 and development. These projects are collectively part of the National Geothermal Data System (NGDS), a distributed, online networked system for maintaining, sharing, and accessing data in an effort to lower the levelized cost of electricity (LCOE). Most of the projects involved in the NGDS conclude on or before April 30, 2014. The exception is the "State Geological Survey Contributions to the NGDS" project which received a no-cost extension through December 31, 2014 to deploy the Node-in-a-Box (NIAB) developed under the Architecture, Design, and Maintenance award to Boise State University. 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.

The NGDS vision is that the functions, accessibility, and availability of data will encourage continued participation within the NGDS, from all system users – data consumers, data providers, and application developers. As each data node adds to its data repositories, the system-wide 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, an example of such is the recent adoption by the State of Arizona to deploy a network built on the NGDS underlying framework for natural resource, environmental, and transportation data. Thus, there are benefits to each participant to continue to add data to the system and maintain it. The long term goal is that the data network reach a 'tipping point' at which it becomes like a data equivalent to the World Wide Web – where everyone will maintain the function because it is expected by its clientele and it fills critical needs. Applying this vision to the NGDS, it also opens the door for additional data providers external to geothermal development, thus increasing the value of NGDS and its underlying data integration platform, USGIN.

NGDS was formally launched by U.S. Secretary of Energy Dr. Ernest Moniz on May 28, 2014 at the White House Energy Datapalooza in Washington, DC, USA.

2. THE NATIONAL GEOTHERMAL DATA SYSTEM (NGDS) VISION

2.1 Goal

The ultimate goal of the National Geothermal Data System (NGDS) is to support the discovery and generation of geothermal sources of energy. The NGDS will provide online access to important geothermal-related data from a network of data providers in order to:

- Increase the efficiency of exploration, development and usage of geothermal energy by providing a basis for financial risk analysis of potential sites.
- Assist state and federal agencies in making land and resource management assessments.
- Foster the discovery of new geothermal resources by supporting ongoing and future geothermal-related research.
- Increase industry and public awareness of geothermal energy development potential.

2.2 User Communities

The National Geothermal Data System is a network of three linked communities:

- Data providers who will expose information to the system through standardized, internet-accessible interfaces and interchange formats.

- Software developers who will build applications that utilize the data in the system, and make it easier for end-users to interact with the system.
- End-users who will utilize the software and information provided by the system in order to understand and develop geothermal resources.

2.3 System Architecture

The NGDS includes data covering a wide range of topics, from well logs and drilling data to temperature, geochemical, and geophysical measurements. Standardized data access to important datasets will facilitate utilization of these information resources.

A key component of the system is the catalog service through which data providers register the availability of resources, and through which users discover, evaluate and access resources. A resource is considered part of the system when it can be located by searching the catalog service, which returns a metadata record describing how the resource can be accessed. Data providers maintain data repositories, or 'nodes' in the network, connected through the use of standardized metadata for describing resources, content models for geothermal data, and common web-service protocols for exchanging information. These standards were developed in conjunction with the US Geoscience Information Network (USGIN), thereby providing interoperability with a wider range of geoscientific information.

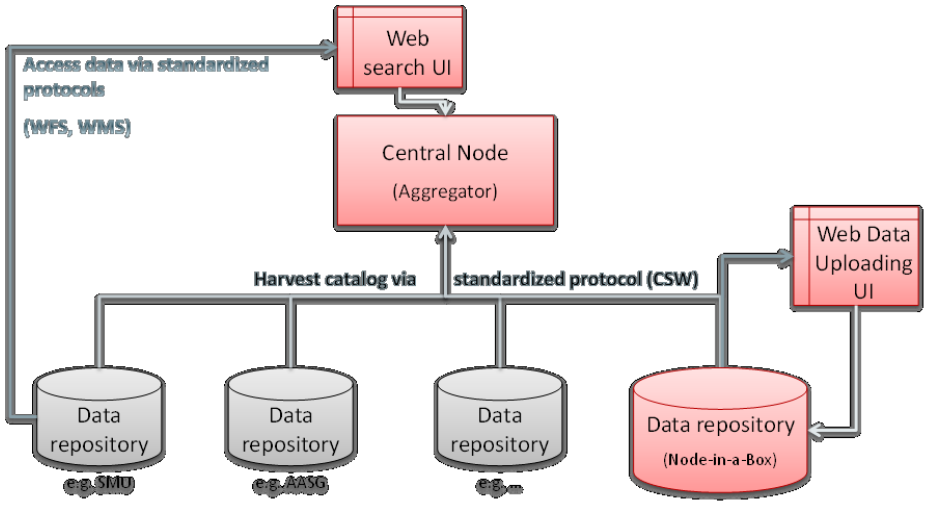

Figure 1: NGDS System Architecture, where data repositories or nodes are fed into the larger system through use of standardized protocols.

2.4 Data Provider Software Package

A redistributable, open-source software package has been created to give data providers a simple way to register data sources, load data and expose those data as a node in the NGDS. The software supports batch import and upload of shared datasets in supported formats adhering to standard content models. The use of this software is not required in order to participate as a node in the network, but is provided as an easy system entry point as well as a beginners data management tool. Data providers may use whatever tools they wish to expose their data, as long as they utilize interchange formats and web-service protocols conforming to NGDS specifications.

2.5 End User Software

End-users may interact with the system through a variety of entry points, but the project implemented two primary access points. As much as possible, these will be integrated in order to appear to the end-user as a single web-based experience. The two primary access points are as follows:

2.5.1 NGDS Website

The website is designed to provide information about the NGDS. It serves as an entry point to the system, allowing users to discover data and applications that utilize NGDS resources. The site includes information on the project's progress, NGDS specifications, access to the Map-Centric Search Application described below as well as other software applications utilizing NGDS data, presentations, documentation and tutorials, a catalog of NGDS nodes, and any other results as they become available.

2.5.2 Map-Centric Search Application

A user-friendly, web-based application has been created to support finding, visualizing, mapping, and acquisition of data by end-users. This application allows users to discover and access resources made available across all NGDS nodes, and to search for data across the system based on topic, location, time, provider, or key words. Standardized metadata describing each dataset provide the user with the information necessary to determine the utility of that dataset for their purposes. Geographic datasets visualized through a map interface also allow users to inspect the details of individual data points (e.g. wells, temperature measurements, etc.) from properly formatted datasets. In addition to visualization within the application, the interface provides the information necessary for users to access the data from other, third party software applications.

3. US GEOSCIENCE INFORMATION NETWORK

The NGDS is designed based on a service-oriented approach using open standards to support data access by a wide variety of software applications, promote novel approaches to data analysis, and foster the development of tools by third parties. NGDS is based on the U.S. Geoscience Information Network (USGIN) which enables users to efficiently find, access, and share geoscience data, reducing the time and effort spent locating and integrating useful information and document new data by providing for information registration (by providers) and discovery (by users) based on standardized catalog services and metadata. The system accommodates resources in various forms, from unstructured text and images to documented, community Web services and interchange formats. To simplify, USGIN is a collection of Web-accessible resources that are registered in online catalogs and conform to data-sharing practices. Developer specifications for USGIN are available via the USGIN GitHub site at https://github.com/usgin.

The core components of the network are information exchange specifications and the catalog function. Information exchange specifications are community agreements on the conventions necessary for the interoperable exchange of some particular information. By using an interchange exchange it allows data in a variety of formats, organization, and structure to be integrated without having to manually transform it. The catalog function comprises a collection of metadata records that describe resources accessible through the network, and a special information exchange that defines metadata content, how the metadata collection is searched, and how metadata are encoded in search responses.

As a system, USGIN is a loosely coupled system of independent data providers, client applications, and infrastructure. The infrastructure includes 1) tools for registration of new resources, searching metadata catalogs, authentication, and resource validation; 2) registries for vocabularies, agents, specifications, and interchange schema; and 3) documentation and educational resources. Because network operation is based on information exchange specifications that are independent of any particular hardware or software implementation, all of the operational components can evolve as technology evolves. Use of standard protocols enables data access using off-the-shelf software, both commercial and open-source.

The distributed nature of the system means that stewardship of resources is determined by the resource owner. Participation in the network requires that a resource provider create metadata that conforms to the profile, and make the metadata and the described resource available. The network is open; anyone can deploy new nodes and components that implement one or more USGIN specifications, without requiring approval. New specifications can be introduced for service protocols, interchange formats, or vocabularies. Similarly, the system back end is decoupled from the front end, meaning anyone conforming to system configuration can build a portal into the system, consuming the data feeds and adding their own applications, visualizations, or other value-added functions. The International Renewable Energy Agency is already taking advantage of this, consuming key data sets for inclusion in the Global Atlas of Renewable Energy, accessible via the IRENA website www.irena.org. Keeping resources under the stewardship of the parties responsible for the information promotes system sustainability because the stewards have a direct connection with the quality of the product rather than submitting the data to a centralized database and related data manager.

3.1 NGDS Data Provider Software Package and End User Software

While USGIN permits application of a variety of service protocols, to simplify participation by NGDS data providers a simplified software stack, Node-in-a-Box (NIAB) was developed under the NGDS Architecture, Design, and Testing Project (NGDS Design Project). The NIAB software targets organizations or individuals who wish to host at least one of the following:

- An online repository containing resources for the NGDS.
- An online site for creating metadata to register resources with the NGDS.
- NGDS-conformant Web APIs (application programming interface) that enable access to NGDS data (WMS, WFS, WCS).
- NGDS-Conformant Web APIs that support the discovery of NGDS resources via catalog service (CSW).
- A web site that supports discovery and understanding of NGDS resources.

While a number of different open-source frameworks were identified for the development of this software stack, including existing systems in use by the NGDS, the NDGS Design Project determined to use the Open Knowledge Foundation's CKAN (http://ckan.org) implementation as it provided the closest match between out-of-the-box functionality and the NGDS NIAB requirements (see https://github.com/ngds/documents).

A major consideration in the evaluation of the software system was the long-term viability (sustainability) of the eventual application. By building on an existing, active and widely used open-source project, including by the US Office of Management and Budget for use on http://data.gov, it ensures that external developers can assist in the ongoing maintenance of the system. Other factors considered in the development framework included; 1) adaptability of the user interface to be compatible with an independently developed user experience concept for NGDS users; 2) ease of extensibility, with a plug-in architecture that allows addition of functionality without having to modify the core codebase; 3) support for geographic data and map-based search and data browsing; 4) support for administrative activities like user management, access control, and activity logging.

CKAN is written in Python and makes use of a variety of open source frameworks including Pylons, which itself is a combination of various open source frameworks integrated to form the basis for Web-based Enterprise-level applications. The primary CKAN user scenario is data storage and management and includes file storage, metadata management, and management of structured data. In addition, it offers a plug-in mechanism enabling developers to rapidly extend CKAN's core functionality; most importantly to the NGDS, the ability to support geographic features as well as exposing metadata according to the OGC standard catalog (CSW). CKAN implements crucial housekeeping features such as user management and logging, which can be tedious to implement but crucial for the site's usability.

While there are many attributes to the out-of-the-box CKAN functionality, there are a number of items that required NGDS developer modification including:

- Modification of the User Interface (UI) to reflect the user experience testing as completed through the NGDS Design Project UI testing.
- Creation of the NGDS/USGIN metadata, including spatial extent.
- Consistency check for well-known structured data files (described in the next section).
- Providing OGC services for uploaded structure files with geospatial information.
- Full-text indexing of documents.
- Role-based right for uploading and publishing data.
- User feedback and rating of uploaded data.

The NGDS CKAN code is currently available for deployment at https://github.com/ngds/ckanext-ngds/ as well as registered with the CKAN developer code repository (https://github.com/okfn/ckan). During the sustainability maintenance phase, USGIN developers will interact directly with the Open Knowledge Foundation to provide the following:

- A more robust installation script.
- A mechanism for script updates to installed instances.
- Clean-up unused/redundant code.
- Ongoing organization and cleanup of the install instructions and repository.

Using CKAN as the NIAB solution and primary UI meets many of the needs and software specifications outlined in the NGDS Software Requirements Summary, version 2.7, November 5, 2012 developed by the NGDS Design Project.

3.2 NGDS Software Requirements & Achievements

Based on the Requirements Summary, the following have been achieved and are presented as they have implications on the maintenance and sustainability of the network.

3.2.1 Maintenance:

- All project source code has comments, at least on a per-class level.
- The system's architecture and configuration parameters are well documented.
- The system's source code is covered by unit tests to at least 50% coverage.

3.2.2 Usability and Accessibility:

- The system has a reasonably simple to use installation tool, with detailed instructions that guide the user through the process of installation and joining the network.
- The system is cloud ready, archived, and available via a Virtual Machine using Amazon Web Services.
- The system uses project developed graphical user interfaces for a uniform look-and-feel for web applications.
- The project-developed applications provide online help explaining how to perform user-related functions.
- Key data import operations are transactional, i.e. the user can abort operations before completion without any negative consequences.
- The project has a status indicator for data uploads.
- The system underwent review for compliance with section 508 amendment to the Rehabilitation Act of 1973 and ISO/TS 16071 "Ergonomics of human-system interaction – Guidance on accessibility for human-computer interfaces.".

3.2.3 Performance and Scalability:

- The system was designed to support a maximum of 1,000 concurrent users.
- Each data provider node is capable of maintaining at a list of at least 100 other NGDS nodes for harvest or distributed search.
- Each data provider node is capable of supporting up to 50 simultaneous authenticated, logged-in users, at least 50 HTTP requests every minute, and takes no more than 10 seconds to respond to each request.
- They system can handle the import of data files up to 2BG in size, import up to 1,000 data files in any one import operation, store up to 100,000 data files in the import director of each data provider, and store up to 500GBs of data files (this is completed by the use of Amazon Web Services and thus is expandable to meet the needs of the system).

3.2.4 Security:

- Valid login authentication is required for all data submitters, stewards, and administrator functions.
- Permits only valid users to write data they have permissions to write and delete data they have permissions to delete.
- The data provider node maintains the integrity and availability of all data stored in its local data store and repository.

3.2.5 Supportability:

- NGDS components are written using standard coding styles for the programming languages used (Python, Java).
- NGDS software is designed utilizing the concept of encapsulation, that is, components are encapsulated by related functionality.

4. SYSTEM OPERATIONS

4.1 Code Repository

As part of the systems commitment to open source software development two GitHub repositories for system design, specifications, documents, and code were created and are maintained by the Geoinformatics staff at the Arizona Geological Survey. The repositories are open for participation and available for public viewing.

Main NGDS Repository: https://github.com/ngds.

CKAN Development Repository: https://github.com/ngds/ckanext-ngds/.

NIAB Installation Repository: https://github.com/ngds/install-and-run.

Programming languages used for the NGDS CKAN extension build are Python and JavaScript, with OS (open-source) components Apache SOLR and Tomcat, PostGIS, GeoServer, and GDAL. These were selected due to their support and use in the open source community and have strong communities of practice surrounding their use. Likelihood that the languages and components will be obsolete in the near future is slim.

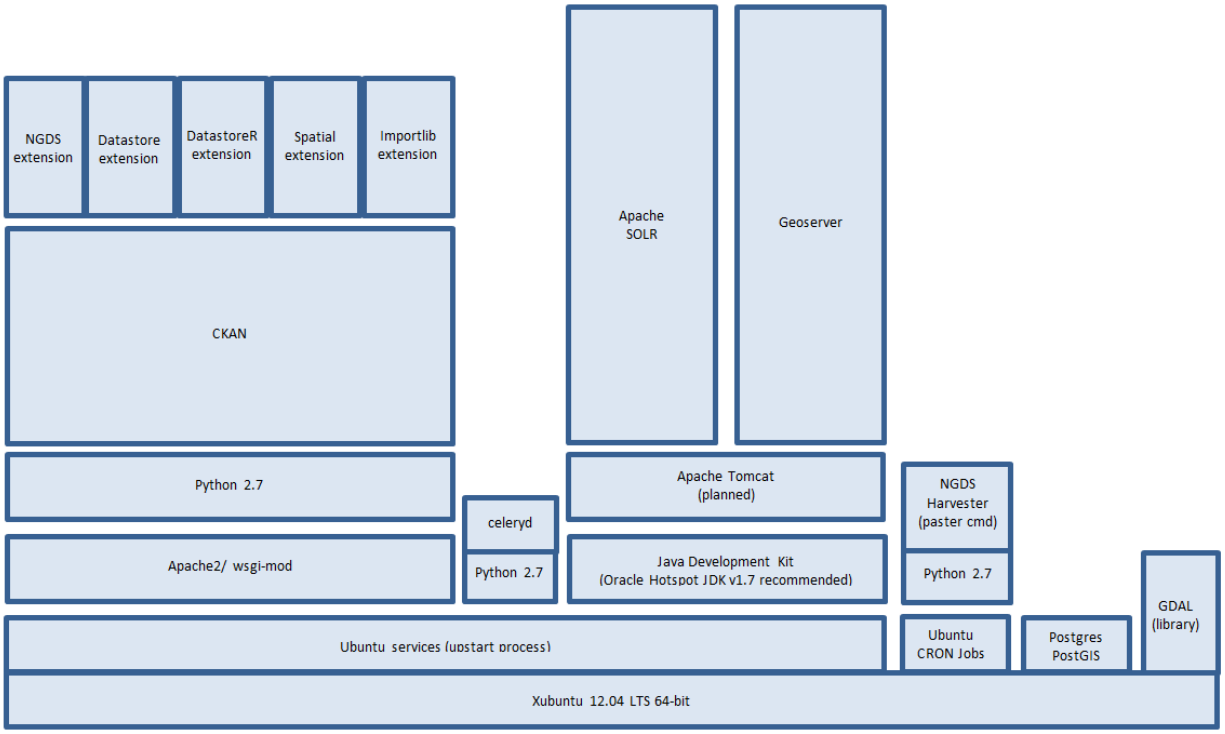

Figure 2: Schematic of components in NGDS CKAN extension software stack.

4.2 Node In A Box (NIAB) Code

This section is specifically on the maintenance of the NIAB code and does not include details on the data maintenance and installed nodes. For information on data maintenance and installed nodes, please view the next section "Maintaining Content.".

AZGS is currently in the process of working directly with the Open Knowledge Foundation to increase the viability and usability of the NGDS NIAB as well as to better align the NGDS extensions to the CKAN project code. This work will occur from May 2014 to December of 2014 to help improve the installation of at least six NIAB at State Geological Survey's across the country (additional information provided in the Maintaining Content section). This work will also make additional developers aware of the NGDS extensions, adding developers to the code team, and thus the maintenance of the extension code.

The current version of the code is tagged at version 1.0.1 and available at https://github.com/ngds/install-and-run. As improvements are made and new code is pushed to the GitHub master branch, newer versions will always be tagged. Current users will be alerted by RSS feed, email, or other notification system when a new version is available, along with a list of upgrades and changes indicated for that version.

4.3 Licensing

Per the requirements of the NGDS Design Project, the software produced would have an open source license variant allowing users to copy, distribute, and transmit the software; adapt software for other applications; and make commercial use of the software under the condition that the attribution of the software is included in any copy or derived work. Thus, NGDS selected the GNU Afferno General Public License v.3 and the Creative Commons Attribution 3.0 Unported License. All NGDS software and development materials are licensed in this capacity, including the website and site materials.

- GNU Afferno General Public License, v.3 – specifically for software designed to ensure cooperation with the community, ensuring that users can share and change all versions of a program while maintaining its free use in the community.
- Creative Commons Attribution 3.0 Unported License – requires that users provide appropriate credit, provide a link to the license and indicate if changes were made when they share (copy and redistribute the material) or adapt (remix, transform, and build upon the material).

Currently NGDS does not track or monitor usage or popularity of search terms due to the intellectual property associated with exploration activity. The NGDS website at http://geothermaldata.org does monitor page visits, DNS locations, and response time of the site to assist with improvements of the site.

In addition to the licenses for NGDS materials, the website is dedicated to the public domain through the Creative Commons 0 1.0 Universal (CC0 1.0) Public Domain Dedication. Thus the http://geothermaldata.org site includes the following statement:

The content of this web site is dedicated to the public domain. All rights to the work are waived worldwide under copyright law, including all related and neighboring rights, to the extent allowed by law.

The NGDS requests that it be acknowledged as the source in any subsequent use of its information. Some materials on this site have been contributed by private individuals, companies, or organizations and include a copyright notice. It is the user's responsibility to contact copyright owners and obtain the written permission required under U.S. copyright law before using these materials.

Links may be made to the NGDS Web site from personal and organization Web pages. NGDS requests that you link to its site rather than downloading portions of the site to another Web server so viewers will see the most up-to-date information.

NGDS materials may not be used to state or imply the endorsement of DOE EERE, AZGS or any of their employees of a commercial product, service, or activity, or be used in any other manner that might mislead the public.

4.4 Website and User Portal

The site http://geothermaldata.org is the primary site that provides information about the NGDS to all participants as well as serves as the gateway to the system to discover data and applications that utilize NGDS resources. In addition the site provides informational resources on the project's progress, access to NGDS specifications (on the GitHub), access to a map-centric and library style search interface, access to other software applications utilizing NGDS services (Resources), NGDS tutorials (via YouTube and USGIN site), and a link to the catalog of NGDS nodes.

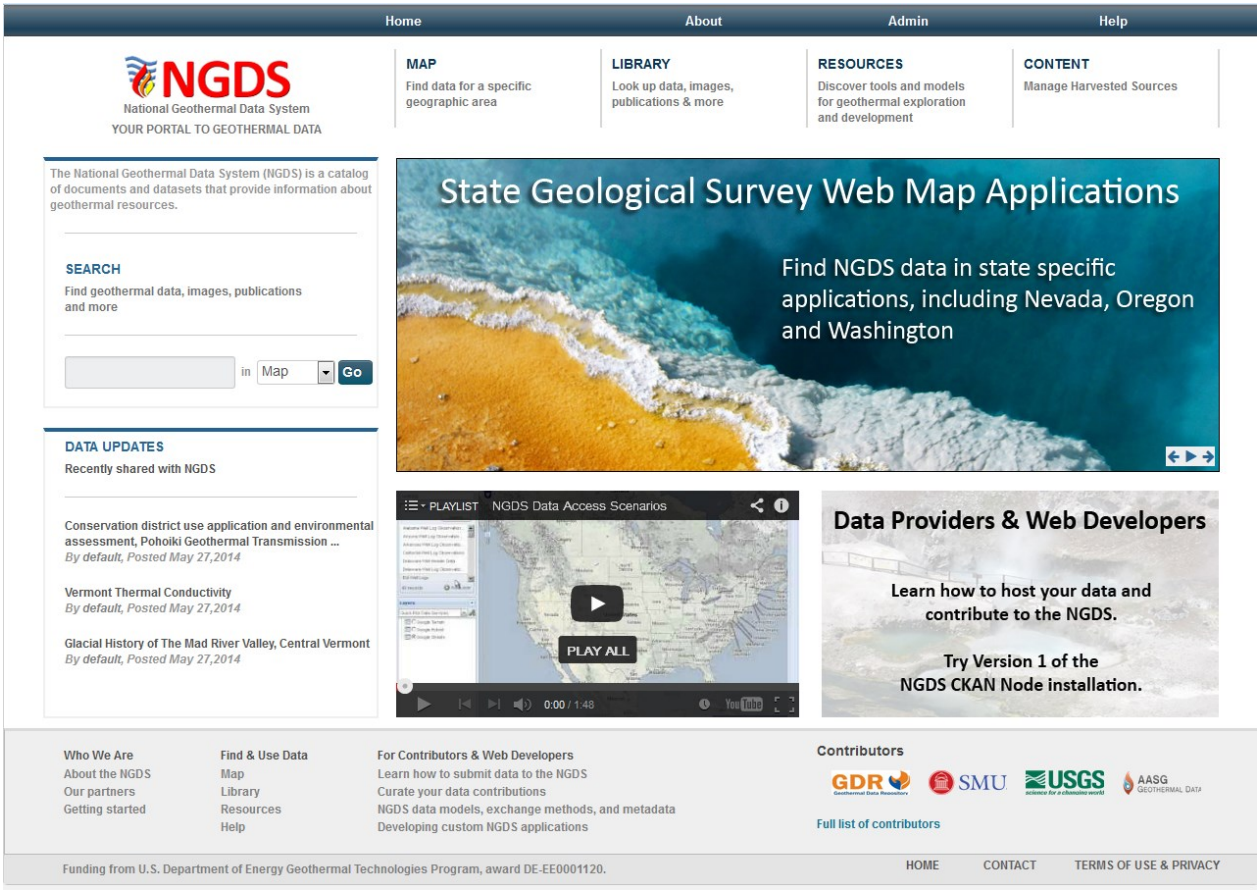

Figure 3: Current homepage of NGDS, with rolling screen of featured data providers, how-to videos, and other vital information about the project.

While the primary site is http://geothermaldata.org, we have also procured the following domain names and provided redirects to the primary site:

- Geothermaldatasystem.org.
- Usgeothermaldatasystem.org.

In addition, we have the http://stategeothermaldata.org and http://usgin.org domain names (as well as all associated subdomains of the five total URLs). All domain names are registered through the year 2020 to provide consistency in URLs. Currently all domains are hosted via Site5 on a shared server with unlimited disk space, bandwidth, and websites. However, due to recent issues in support with Site5 it was decided that the NGDS and related sites would move to KnownHost which guarantees an independent Virtual Private Server, thus providing additional power and speed to the sites. The system at KnownHost provides 1536MB of guaranteed RAM, 60GB of disk space, and 4000GB of Bandwidth for approximately $40 per month. This should accommodate the NGDS and related sites well into the future. However, should the sites require additional RAM, disk space, or bandwidth, there are five additional options for hosting at varying costs. Service includes a 99.9% SLA, free backups, high performance VPS hosting, and 24/7 support.

Additional advantages to managed hosting includes no hardware requirements, all downtime is covered by the hosting company, and the information/sites are easily transferrable should a new entity require management of the sites. Finally, the new hosting service provides a full dashboard for site statistics and monitoring.

In addition to general website hosting and maintenance, the NGDS requires cloud services for the NIAB GeoServer and multitude of Web Services. Throughout the duration of this project we have used Amazon Web Services (AWS) to meet these needs. AWS costs are established via the traffic and requested redundancy. Currently NGDS services cost approximately $150 per month. While we expect traffic to increase due to the v1 release of NIAB, we anticipate costs of approximately $200 per month for AWS for the near future (12-18 months). During the maintenance phase through April 30, 2015 we will evaluate traffic and adjust this plan accordingly.

4.5 Service Notifications

NGDS nodes and hubs are encouraged to use the USGIN RSS feed for service notifications when deploying web services. The tool is available at: https://github.com/usgin/service-notifications. This permits data providers the ability to quickly distribute notifications regarding the services that they host, such as planned downtime, unforeseen server issues, new service availability, etc. Subscription to the service notification is available via the USGIN notifications RSS feed on the GitHub.

5. MAINTAINING CONTENT: DATA SERVICES, APPLICATIONS

5.1 Regional Server Hubs

Four NGDS Regional Server Hubs (housed at state geological surveys in IL, KY, AZ, and NV) have been thus far tasked with hosting web services for their surrounding data providers (other state surveys or institutions) who lacked the technical capability to host their own data in web service format. Deployment of the Server Hubs includes limited troubleshooting time, monthly data updates, backup and disaster recovery costs, and institutional facilities and administration costs. Currently 3 of the 4 hubs are using shared ArcGIS servers (1 has a dedicated ArcGIS server) and independent institutional infrastructure, as the concept of USGIN enables participants to use existing infrastructure systems with specified protocols and standards.

NGDS tier 3 data services are discrete data types where aspects are controlled by static schemas. One aspect that is specified in the schema is the layer name of a given data type; this assists with search and access of the data. As the hubs host multiple tier 3 services from multiple states, they therefore serve multiple services of the same data type. This becomes an issue with the current CKAN NIAB application.

The current construct of the CKAN NIAB application is problematic for hubs or other entities who might serve multiple tier 3 web services of the same type. This is due to the GeoServer web server platform used in NIAB where only one layer name can exist in any given GeoServer instance. Thus, even though NGDS hubs could use NIAB for service deployment and management, there are undesirable consequences. First, the web services may become too large for the current processing capabilities of most desktop/laptop machines, and second, the data management scheme (each state having its own dedicated service per data type) and proven successful workflow of the hubs would need to change. Over the life of the project, the hubs have been functioning through use of individualized workflows, which usually include ArcGIS Server and PostGIS for web services management coupled with metadata management at the state geothermal data repository (http://repository.stategeothermaldata.org/repository/).

Considering the possible NIAB limitations for those entities serving multiple datasets of the same type, two options are recommended for hubs moving forward; 1) continued use of the state data repository for metadata management and ArcGIS Server for web service deployment; 2) continued use of the current ArcGIS Server workflow for web service deployment coupled with NIAB metadata management.

Scenario 1: Hubs continue to use ArcGIS Server; Hubs continue to use the UI at repository.stategeothermaldata.org.

Currently, NGDS hubs use ArcGIS software to deploy web services and manage the associated data formats (geodatabases, feature classes, data tables). When project deliverables (datasets) are deployed as web services, a metadata entry is made at repository.stategeothermaldata.org for that dataset with distributions that include the links to the live web services (WMS, WFS, ESRI rest page).

The current workflow of AASG deliverables (data) processing includes the creation of metadata for the services, where a ‘New Resource’ is created at repository.stategeothermaldata.org. Moving forward with this scenario, hubs will be responsible for creating metadata at repository.stategoethermaldata.org, which will require minimal additional training for hubs as they are already familiar with the interface. The biggest hurdle will be to assist the hubs in properly integrating the metadata entry point into their current workflow.

Scenario 2: Hubs continue to use ArcGIS Server; Hubs use NIAB for creation and management of metadata for services, but not service deployment.

As creating Tier 3 web services using NIAB becomes problematic for hubs, this scenario proposes that services are deployed as usual with ArcGIS Server. Instead of entering the metadata for the services at repository.stategeothermaldata.org, the hubs would use NIAB simply for the creation and management of that metadata.

When a "Link to a data service" is added as a resource in NIAB (creating a metadata record), it is possible to add multiple service distributions. These distributions are the WMS/WFS links. Once a metadata record is created for a service in NIAB, locate the Edit button on the page of the resource. Here, go through the "Link to a data service" workflow once more to add an additional link. This will be added to the same metadata record. The hurdle here is much the same as Scenario 1; hubs will decide which methods of metadata management best fit their workflow.

Until which time the NIAB might include the capacity to handle multiple layer names or GeoServer instances, we recommend that the hubs maintain Scenario 1 as installing the NIAB simply for metadata management would cause unnecessary costs for the regional hubs (for additional cloud services and server maintenance on services that they will not be implementing, i.e. GeoServer). This scenario is an acceptable solution though, as innumerable options exist for manageing nodes or hubs in the system if outlined protocols are followed.

5.2 Regional Server Hardware

While the USGIN system is designed to accommodate a variety of data sharing "nodes" and platforms, due to the importance of the multi-state hosting of the Regional Server Hubs, we are recommending a more standardized platform for the hubs while moving forward.

In order to increase the Regional Network Hubs reliability, maximize uptime, and ongoing sustainability we recommend that the Hubs virtual machine host servers be dedicated services rather than using a shared server solution (as is currently the case). We estimate that a Dell tower server in the $2,000 range would be powerful enough to support this type of solution. We also recommend that each of the Hubs use a USB hard drive ($150) attached to the server as a backup drive. The free Windows Backup software included with Windows Server 2012 has the ability to perform scheduled backups to a USB drive that includes a complete image of the entire Hyper-V host server including any virtual machines on the server.

We also recommend setting up Hyper-V Replication (next section) between the hubs Hyper-V host servers for redundancy. If a hub server or virtual machine goes offline, Hyper-V Replication would provide a copy of the hub that could be quickly booted up on the backup Hub server until the main Hub server problem is resolved.

5.3 Replication and Disaster Recovery

AZGS is assisting with the replication and disaster recovery for the other hubs. Delays in virtualizing, due to institutional requirements, have postponed full implementation, although as of March of 2014 each hub was running a virtual machine. Hyper-V is a virtualization solution from Microsoft designed to create and operate virtual machines. Hyper-V replication is a feature of Hyper-V that provides the ability to make a copy of a virtual machine at a different geographic location for disaster recovery purposes. In 2013, AZGS installed Hyper-V 2012 on an Arizona Geological Survey server to support the hubs.

In September of 2013, the NGDS virtual server at the Nevada Bureau of Mines and Geology (NBMG) failed catastrophically with a loss of 200GB of data. Fortunately, Hyper-V replication was engaged and the server mirrored, insuring that all NBMG data survived. Using the Hyper-V mirror to restore 100% of the data avoided a catastrophic loss of data that would have taken hundreds of hours to reconstitute; in this case it took approximately 3 days. AZGS has published a guide for installing and creating a Hyper-V replication environment. It is available via the AZGS repository at http://repository.azgs.az.gov/uri_gin/azgs/dlio/1567.

5.4 Nodes on the Network

An NGDS node is a web-accessible server that hosts at least one of the functional capabilities enumerated below to play a role in making geothermal-relevant data assets accessible.

To be considered a node in the NGDS, the following criteria must be met: The capabilities offered must play a role in making geothermal-relevant data assets accessible; The capabilities offered must be publicly Web-accessible; A node must offer at least one of the following capabilities:

- Host a Web-accessible repository of geothermal-relevant data assets with metadata conforming to the USGIN ISO profile (USGIN Standards and Protocols Drafting Team, 2010-11) published through an NGDS catalog node.
- Host a web-accessible folder that 1) contains NGDS-conformant metadata files and 2) is registered for harvesting by an NGDS catalog node.
- Host NGDS-conformant web services (WMS, WFS, WCS, etc.) that are registered in an NGDS catalog.
- Host an NGDS catalog node, which is a server operating a CSW 2.0.2 service that offers metadata conforming to the USGIN ISO profile.

NGDS nodes should self-identify by providing an NGDS node self-description document. Additional information on Node Specifications are available at: https://github.com/ngds/install-and-run/tree/master/Becoming%20an%20NGDS%20Node

There are currently two types of NGDS nodes; the NIAB installs and organizations that use their in house systems to meet the requirements above. Currently, the following nodes are hosting their own data via NGDS Tier 3 web services:

- Southern Methodist University's Geothermal Laboratory (independent project with independent sustainability plan, SMU has guaranteed to house the web services and data for at least 10 years; how the SMU node will respond to issues with metadata and general server maintenance are unknown)

- Illinois Geological Survey (also a hub).
- Kansas Geological Survey.
- Kentucky Geological Survey (also a hub).
- Minnesota Geological Survey.
- New Hampshire Geological Survey.
- Nevada Bureau of Mines and Geology (also a hub).
- Oregon Department of Geology and Mineral Industries.
- Washington Department of Natural Resources, Geology & Earth Sciences.
- Wisconsin Geological Survey.
- Arizona Geological Survey (also a hub).
- Minnesota Geological Survey.

As part of the NGDS Design project additional installations of the NIAB were to be deployed, including the metadata aggregator node. Currently the metadata aggregator node is housed at the Arizona Geological Survey (AZGS). Three instances of the NIAB are successfully installed at the Arkansas Geological Survey, the New Mexico Bureau of Geology and Mineral Resources, and the University of Utah's Earth and Geosciences Institute. It is expected that as many as eight additional state surveys will install NIAB for data and web service management by the end of 2014.

5.5 Possible Extensions & Improvements on the NIAB

The software specifications and requirements document outline a series of requirements that were not met for the Version 1 release. For example, the bulk upload of data sets was a target that was not met. In addition, testing of the v1 release over the next 6-12 months will generate a series of additional user-inspired requirements that will require implementation. One such example is for a metadata repository extension of the NIAB, that is, a NIAB that does not require installation of the GeoServer platform and can be used by entities or organizations that simply want a mechanism for contributing metadata to the network while hosting their own services. As these requirements are developed, issues will be logged and prioritized on the NGDS GitHub. The AZGS and USGIN Foundation Inc. team will accommodate as many of these priorities as possible.

5.6 Content Model (Information Exchange) Maintenance & Sustainability

Content models, or information exchanges, in use for the NGDS were established using a community developed mechanism and approved by the Department of Energy's Geothermal Data System Development and Population Working Group (GDSDPWG), which was comprised of the technical leads from each project and a series of technical monitors. When available, existing exchanges were used. Information in deprecated exchanges were generally synthesized with another exchange and thus, are still available. All available exchanges are listed at http://schemas.usgin.org/models/.

Should the likely event occur that additional exchanges are required to expand the usage and applicability of the system, the following methods are proposed. Full details are available at https://github.com/usgin-models.

The decision to define a new information exchange should be based on the likelihood that others will want to publish similar datasets in the future. Members of the USGIN community propose specifications for data sharing exchanges. Exchange documents are developed and reviewed using a publicly accessible repository on GitHub (https://github.com/usgin-models). Each exchange has a separate repository associated with the USGIN-models pseudo organization. A proposed model must have an identified steward, and a working group of at least three participants with relevant domain knowledge and understanding of the interchange technology. There is no formal process for defining workgroup membership; normally the challenge is finding a sufficient number of qualified individuals to provide meaningful reviews and comment.

The exchange steward is responsible for assembling the workgroup and assuring sufficient expertise in the group to generate a sound content model and implementation. The exchange steward requests creation of a new model repository at the USGIN-models GitHub from the organization members (currently primarily AZGS Geoinformatics staff), and identifies workgroup members who will have commit privileges on the repository. Any community member can create a repository branch to propose changes using standard GitHub procedures, and request consideration for merging back into the developing model.

After review and approval by the workgroup, a call goes out to a USGIN technical review e-mail list or by RSS feed (http://notifications.usgin.org/) for comments from the community. An open review period of 4 weeks is normal, after which any comments from the community must be resolved to the satisfaction of the commenter. When issues are resolved to the satisfaction of the stakeholders (workgroup and engaged community), the exchange specification is adopted.

When a specification is adopted, all associated documents are copied to a 'tag' branch in the GitHub repository and are not changed after they are 'tagged'. The Specification documents are also copied to the exchange repository at http://schemas.usgin.org, which is a web site set up to provide public access to exchange specifications and any related xml schema documents or other artifacts required for the deployment and operation of the information exchange.

This mechanism ensures that the development of specifications continues to be community led as well as providing a source-trail for the development of such exchanges.

5.7 Maintaining the Existing Data

Experience has dictated that server outages can and will occur. In addition, the potential loss of data from a distributed network, while not as great as "closing the doors" on a traditional database, may occur. In order to preserve the vast amount of data digitization and preservation, the Arizona Geological Survey plans to conduct a one-time data harvest to capture the data available within the NGDS and place it as a back-up and disaster recovery mechanism on both cloud storage and external hard drive. This

process will be complete by May 30, 2014. The external hard-drive will be submitted to DOE as a deliverable under the State Geological Survey Contributions project (end date 12/2014).

5.8 Maintaining the Client Developed Applications

NGDS and USGIN support client side developed applications similar to applications for the iPhone or Android phones. As such, externally developed applications are subject to the maintenance schedules of the developers and not the NGDS or USGIN.

6. ORGANIZATION AND MANAGEMENT

6.1 Non-Profit Organization

We determined that a stand along organization offered the greatest independence and likely adherence to the organizations vision, goals, and missions and thus the longevity of a system such as the NGDS. In the fall of 2013, the Arizona Geological Survey, with the guidance of the Arizona Attorney General appointed counsel, have successfully incorporated the USGIN Foundation, Inc. with the State of Arizona's Corporation Commission as well as filed initial requests with the Internal Revenue Service. Not only is this step in alignment with the recommendations of the May 2013 paper, but also aligns with the Hutchison and Richard 2011 "Recommendations for the Future of the U.S. Geoscience Information Network," joint report to the Association of American State Geologists and U.S. Geological Survey.

The Foundation provides the following expertise:

- Earth Science Data.
- Project Management Excellence & Streamlining of Workflows.
- Managing Geospatial Data.
- Knowledge of Unstructured AND Structured data formats.
- Interoperable Systems.
- Data Provider Community Building.
- Visions for Long-Term Sustainability.

Duties of the USGIN Foundation Inc. will include:

- Management of the Foundation: Strategic Planning; Fiduciary.
- Client Systems Development & Maintenance: Programming/Development; Data Curation; Metadata Maintenance and Curation; Systems Architecture; Systems Administration.
- Business Development, including Outreach & Training: Contract Development and Negotiation; Government and Foundation Grant Development and Negotiation.

7. TECHNOLOGY TRANSFER (EDUCATION, OUTREACH, TRAINING)

7.1 Foundation Website

As part of our improved outreach and future business development we have created a more user friendly website to introduce potential clients to the system. The revised page is at the existing http://usgin.org page with the technical specifications and tutorials moving to a subdomain at http://tech.usgin.org. All product descriptions, once finalized, will be available and accessible via the USGIN homepage.

In addition to the traditional website at http://usgin.org, specifications are available via http://github.org/usgin, and scenario and user video tutorials at www.youtube.com/geothermaldata. The videos available on the YouTube site include 13 data access scenarios guiding users through the search interface as well as a variety of software clients to show the diversity of the system.

7.2 Products & Services

In cooperation with our marketing and product development firm, Blue Canoe Marketing, we have four deliverables, in varying stages of completion, outlining next steps for business development, outreach, and training. These include: 1) the USGIN Suite of Solutions for Open Data Access, 2) a Compliance Guide for Federal Open Data Compliance, 3) the NIAB Marketing Plan, and 4) a 12 month marketing and communications plan for business development.

7.3 USGIN Suite of Solutions

There three types of services that the Foundation will offer: Marketing/Pre-Consultation, Consulting Services, and Support Services. Brief information on each of the types of services is offered below. Note that the services are designed to build upon each other.

7.3.1 Marketing/Pre-Consultation

- Workshops.
- Training & Education.

7.3.2 Consulting Services

- Assessments – provides a plan and recommendations for Open Data Compliance.
- Open Data Strategy and Action Plan – provides a scope of work, specifications, and a proposal to build a basic data sharing system.
- Interoperable Data Strategy and Action Plan - provides a scope of work, specifications, and a proposal to build an advanced interoperable data sharing system.

- Deploying Your Own Node – provides individualized data node with documentation (installation guide, technical support, help desk requirements) and independent data control for the host.
- Hosting – provides a migration plan for sustainability of the independent node.

7.3.3 Support Services

- Open Data Development – provides hands-on modification of data, catalog creation, and customized workflows (based on the results of the Assessment).
- Interoperable Data Network Development - provides hands-on modification of data, catalog creation, customized workflows (based on the results of the Assessment), community building efforts (data sharing affiliates), and information exchange specifications.
- Annual Subscription & Support – provides hosting services, data maintenance, replication services.

Fees associated with each service will depend on the scope and size of the proposed project. These services, while intended to be stand-along services offered to potential clients, can also be revised to respond to funding opportunities in government and foundations. For example, the State of Arizona's Natural Resources Review Council recently approved a plan to develop a Natural Resources Decision Support System (NRDSS) to link data, documents, and GIS layers among nine state natural resources, environmental, and transportation agencies, as well as with corresponding agencies, using a USGIN style system. In response, we will be able to provide a best estimate for total cost of such a system using the clearly delineated services defined above. In this instance, we would include services for assessment of available data resources, and then provide development of the interoperable data network, concluding with a set of proposed annual subscription fees for maintenance and support.

7.4 Compliance Guide for Federal Open Data Compliance

To aid and assist with marketing of a USGIN style network, and to communicate the benefits of a USGIN style network to non-data or technical stakeholders, we have developed an Open Data Compliance Guide in response to the Executive Order 13642 and OMB Memorandum M-13-13. The Guide provides descriptions on the open source tools and resources available for potential client applications that go beyond the OMB requirements to create an interoperable, federated data network, similar to the NGDS. Details described within the guide discuss the benefits of USGIN and NGDS, information on how to create and maintain an Enterprise Data Inventory, how to achieve interoperability with that inventory, and how to create and maintain a public data listing (or USGIN style catalog). In addition, the Guide describes mechanisms for engaging with stakeholders and how to document data that is currently under moratorium. Once completed, this guide will also provide a brief overview of how USGIN can help provide these services.

8. SUMMARY

The central idea of the data access architecture proposed here is that data providers and client applications should be linked through open source interfaces that decouple clients and servers such that they can evolve independently without breaking the system. The hypertext transfer pro-tocol (http) and hypertext markup language (html) are the established protocols and interchange formats in use on the internet, and in the near term these will probably continue to be the main-stay of most interaction in the NGDS.

The OpenGeospatial Consortium Catalog Service for the Web (CSW), currently at version 2.0.2 is the required catalog search and discovery service. The lowest common denominator metadata interchange format using this service to achieve interoperability between metadata provided by various servers is outlined at http://schemas.usgin.org/models/#Metadata and https://github.com/usgin-models/MetadataCompilation, which uses the USGIN profile for ISO metadata.

9. ACKNOWLEDGEMENTS AND DISCLAIMER

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 and DE-EE0002850.

REFERENCES

Allison, M.L. Richard, S., Caudill Daugherty, C., and Patten, K: National Geothermal Data System Sustainability.

Hutchison, V, and Richard, S: Recommendations for the Future of the U.S. Geoscience Information Network, *Joint report of the Association of American State Geologists and the U.S. Geological Survey [Unpublished]*, 11p (2011).

Plan, v1.0: *Arizona Geological Survey*, (2014), 20p, github.com/ngds/documents/blob/master/NGDS-Sustainability-Plan-v1.docx.

System Design Framework for National Geothermal Data System, version 1.0, *USGIN development team*, (May 1, 2014), github.com/ngds/documents/blob/master/SystemOverview_v1.docx.