System Design Framework for National Geothermal Data System

USGIN development team, S.M Richard, editor

2014-04-29

# Introduction

This report presents a framework for National Geothermal Data System (NGDS) architecture focused on the network data publication and data access aspects of the system. The design presented here is based on work done at the Arizona Geological Survey under auspices of DOE award DE-EE0001120, in conjunction with related work on development of the Geoscience Information Network (US GIN) supported by NSF grant EAR-0753154, and a parallel DOE award DE-EE1002850 to compile and publish geothermal data from state geological surveys to integrate with the NGDS.

The design is based on development work over the last several years under the auspices of a variety of geoinformatics projects, and is intended to provide an incremental development framework that utilizes existing technology wherever possible, builds on a variety of existing standards and specifications, and allows for agile development of the NGDS in the current, rapidly evolving technology environment. In order to be sustainable, the NGDS design will have to provide a framework to promote community engagement and incorporate new technology and ideas as they are developed without disrupting existing practice.

This document includes an introductory section discussing the scope of the system based on the original Department of Energy Funding Opportunity Announcemnt (FOA), the NGDS consortium’s proposal, and use cases that help scope requirements. The second section outlines the architecture for distributed data access in the system. The third section discusses data acquisition, and a final section consists of some technical discussion and a summary of recommendations.

## Scope and purpose of system

As described in original Department of Energy Funding Opportunity Announcement (FOA):

National Geothermal Database Description

The National Geothermal Database will store critical geothermal site attribute information such as temperature at depth, seismicity/microseismicity, fracture maps, drilling data, permeability data, well logs, geophysical surveys, etc. The database should be inclusive of all types of geothermal resources such as hydrothermal, geopressured, Enhanced Geothermal Systems, geothermal fluids coproduced with oil and/or gas, etc. It should also utilize information from existing USGS geothermal resource assessments and DOE funded R&D projects. This standardized set of geothermal resource data will be made available to the public and serve to focus geothermal exploration activities, thereby mitigating investment risks.

From <http://apps1.eere.energy.gov/geothermal/projects/projects.cfm/ProjectID=27>:

“The NGDS will be able to handle the full range of geoscience and engineering data pertinent to geothermal resources as well as incorporate data from the full suite of geothermal resource types. It will be able to handle data on geothermal site attributes, power plants, environmental factors, policy and procedure data, and institutional barriers. It will provide resource classification and financial risk assessment tools to help encourage the development of more geothermal resources by industry. It will be an easy to use system that meets the needs of the professional and the public for information on geothermal resources.”

Abstracted from Original Project Proposal from the Geothermal Data Coalition:

*Goal: build a state-of-the art data system.*

* *reduce social-cultural barriers that could hinder the development of a comprehensive database*
* *Provide access to critical data and data products.*
* *Provide the basis for financial investment risk analysis.*
* *Provide geothermal-resource information to the public and decision-makers*
* *support state and federal agencies with land and resource management missions*
* *support ongoing and future geothermal-related research*
* *contribute to enhancing the education pipeline for careers in the geothermal energy industry*

## System Technical Design principles

The National Geothermal Data System must provide online resources to make it easy for users to extract, assess, and synthesize data according to criteria they select. Data will be provided by a community of data providers, many of whom maintain their own data management systems. There are also numerous kinds of existing, “legacy” data in various tables, spreadsheets and databases that need to be made accessible through the system, as well as many documents that are or could be in digital form and accessible through the system. Some of these legacy data are ‘orphaned’ in that the original producer of the data is no longer involved, and there is no acting steward for the data.

Resources (e.g. data, metadata, catalogs, services, tools) are made accessible through the system by creating metadata conforming to a shared content model and sharing them through the main metadata aggregator The metadata provide information describing resources that can be indexed for discovery by search engines, information about provenance and quality of the resource so users can evaluate the resource for their application, and information describing how to access the resource. The access instructions should be in a format that can be utilized by software clients to automate the access process and minimize the amount of user interaction required to bring the resource to their desktop.

Metadata from publishing nodes on the NGDS is aggregated and searchable at <http://www.geothermaldata.org>, providing a single entry point for users to search all resources in the system. Any search client that implements the system catalog service profile should be able to conduct search against any system catalog that also implements the profile. This means that there can be multiple portals and client applications for accessing system resources; a single client can search different catalogs in the system without the user having to reconfigure the software.

Structured data are provided through NGDS services that have published protocol and documented interchange formats. The idea is that multiple data providers can present the same kind of information in the same way, and a client that implements an NGDS service can access that service from any server in the system that offers that service and get data that integrate with minimum operator intervention.

The following bullet points are extracted from the original project proposal and subsequent Statement of Project Objectives (SOPO) to help clarify the scope of the project.

* Design must be expansive; capture the full physical, geologic, geophysical, and geochemical context of geothermal systems on scales ranging from regional to the individual well bore to the thin section and microscopic scales.
* Information in system must be supported by metadata to document authority and to provide people and projects that compile data the appropriate level of recognition and support
  + All data will credit the original intellectual source and host server of record for that data.
  + Standard measures of "quality" should be available. E.G. variability, bias, systematic error, imprecision, accuracy, precision, reproducibility, etc.
* Able to adapt to evolving requirements, new technologies and standards, and expanded scope as necessary.
* Use existing or emerging standards and technology whenever possible rather than developing new ones
* Open source and open accessibility is preferred to encourage third parties to independently develop software applications that can use the content and services provided by the system
* People who produce data can integrate those data into the data system.
* Provide a means of capturing legacy data
* Distributed data system, connected by the principle of data sharing and interoperability among linked sites
* Two-way system of both data-in and data-out.
* Provide the users with the base data behind data products
* Assign Digital Object Identifiers (DOI) to datasets
* Accessible through multiple browsers
* Easily maintained

### Data Access

* Provide open access to public data
* Contributors can require user consent to license conditions on data (e.g. noncommercial use only)
* Implement access controls and security to limit access to datasets at discretion of provider
* Data owner retains control of access to all data regardless of where it is stored.

## Approach

One of the basic objectives of the NGDS is to make access to data simpler. A major time consuming aspect of bringing disparate datasets together is data integration. This process involves matching field or element names in the schema for various data sets, selecting those that contain the information of interest, and then merging content into a single data set with consistent usage of vocabulary and units of measure in a standardized collection of fields or elements. Data integration may be done by data providers who choose to deliver data in standardized interchange formats, by data consumers who acquire data in heterogeneous formats and schema and figure out how to extract what they need, or data integration may be done by middleware layers that implement transformations between known formats and schema.

Data integration in our current system of scientific information interchange is mostly left to the data consumer. Until recently, the most common approach was has been for an investigator to collect various datasets and integrate them into a single database that was used for some analysis; some small part of the data might get published, and the compiled dataset was subsequently committed to oblivion. Centralized data aggregation schemes have also been developed and deployed, but rarely outlive project funding or are not maintained and rapidly grow stale due to out-of-date data or use of retired technology. A tremendous amount of effort has been made towards developing systems to promote the management of data such that it may be reused without having to repeat the same integration and cleanup processes over and over.

The path adopted for the Geoscience Information Network to simplify data access and promote reuse is to develop standard formats and access protocols used to deliver common data sets (e.g. borehole temperature data, heat flow measurements) to consumers. The onus of data maintenance is shifted towards organizations that are tasked with data management and preservation. By documenting data schema, encoding formats and practices for vocabulary usage, data can be put into the ‘data integration’ format, or ‘information exchanges’ when it is made available on the web. Because of its enhanced utility in a standardized format, management and preservation of the data are more strongly motivated.

This requires education of the data providers/publishers on the use of the information exchanges, but results in a larger community of IT personnel who know how to get data into and out of the information exchanges. Mapping data into an interchange format is likely to be done more accurately by those who originate the data working in conjunction with data managers who understand the interchange formats. The net effect is a greater likelihood that the federated information system using the documented interchange formats will outlast any particular researcher, data provider, project, or agency. HTML on HTTP, NetCDF, and XML are examples of data integration formats that have achieved wide usage and long term usefulness.

The use of schema and encoding specifically designed for data integration and interchange means data producers and consumers can continue to use internal data formats that are optimized for their business requirements. Use of the community interchange formats reduces the amount of work required because only one transformation from internal to interchange format has to be engineered for each interchange format in use.

Data integration by providers introduces additional costs into the data delivery process, and this cost dictates that there must be consideration of the benefits obtained. For data that are not provided using documented interchange formats, detailed metadata describing the schema and encoding of the data will be necessary to enable reuse. The NGDS steering committees originally developed information exchanges based on input from experts in the geothermal community, but now that process has moved to the broader community. Policies regarding defining new, relevant information exchanges and determining what data should be presented in and in what formats,as well as what data are specialized to a degree that data integration by the providers is not warranted for the broader system are determined by the users of the system. Criteria for such decisions will likely include how many providers have a particular kind of data, how often that kind of data are known or expected to be used, the cost of obtaining or reproducing the data, and the expected useful lifetime of the data.

## Requirements

#### Data discovery

The fundamental use case addressed by a catalog system is to find resources of interest via the internet, based on criteria of topic, place, or time, evaluate resources for an intended purpose, and learn how to access those resources. Detailed metadata describing a resource data schema, describing service or application operation, or providing detailed descriptions of analytical techniques and parameter are outside the scope intended for basic search and discovery metadata. Our contention is that this more domain/resource specific type information is better accounted for with linked documents utilizing schema appropriate to those specific resources. Some examples include OGC getCapabilities, WSDL, and ISO 19110 feature catalogs.

At the main aggregator <http://www.geothermaldata.org>, data discovery is accomplished in several ways:

* The NGDS Map Search, <http://www.geothermaldata.org/ngds/map>, which allows a keyword or drawn bounding box search for GIS data as well as geo-located publications and other resources.
* The NGDS Library Search, <http://www.geothermaldata.org/dataset?_tags_limit=0>, having keyword as well as facted search functions. The faceted searches can be performed using authors, contact, broad categories, and by content model (information exchange types).
  + Search hints, using special terms, wildcard searches, and special characters is available at <http://ngds.github.io/documents/SearchHints.htm>.
* Geothermal Prospector preview, <https://maps-stage.nrel.gov/geothermal-prospector/>, which is provided in the Dataset Details page for each resource that has a WMS and/or WFS distribution.

As discussed earlier, system design dictates that any software client can access the data in the system given the predictable data types and web service distributions. Some of those client-side access options include:

* The Data Exlporer, <http://data.geothermaldatasystem.org/>
* The USGS National Map Viewer, <http://viewer.nationalmap.gov/viewer/>
* Google Earth, <http://www.google.com/earth/>
* Any number of FOSGS (free and open source geospatial software) programs

See <http://www.geothermaldata.org/ngds/resources> for more information on user access.

#### System Architecture Requirements

The system consists of the following components:

* Set of standards, content models and protocols that allow the various components in the system to function interoperably
* Network of distributed nodes that provide geothermal data following the standards and protocols defined by the system
* A variety of applications built to search for, or consume data using the system’s standards and protocols
* Searchable catalog that maintains a registry of data provided by all the nodes in the network
* A mechanism for creating and dereferencing unique identifiers (or URIs) for resources in the system. “Resources” include metadata records, datasets, individual rows within a dataset, and physical objects
* Web-application that serves as an single entry point to the system in order to establish and maintain system identity and branding
* Data in the system is categorized in three distinct groups, or tiers
  + Tier 1: unstructured data such as scanned well logs, geologic maps, physical resources (e.g. core, rock or water samples, etc.), or peer-reviewed scientific articles
  + Tier 2: tabular, structured, georeferenced data that does not conform to a content model
  + Tier 3: tabular, structured, georeferenced data that does conform to a content model
* The system defines a set of content models for common geothermal data themes and provides a canonical encoding of these models
* The system defines a standard model for metadata describing data in the system and provides a canonical encoding of that model
* The system defines a set of protocols used to exchange data, metadata and other information between components of the system
* File-based data is transferred via standard HTTP protocols (e.g. GET), or through person-to-person communication in the case of physical or otherwise offline resources
* Tier 2 data can be transferred via standard Open Geospatial Consortium (OGC) web services (e.g. Web Feature Service or WFS, Web Map Service or WMS, Web Coverage Service or WCS)
* Tier 3 data are expected to be transferred via standard Open Geospatial Consortium (OGC) web services (e.g. Web Feature Service or WFS, Web Map Service or WMS, Web Coverage Service or WCS)
* Metadata is transferred via a standard OGC web service, Catalog Service for the Web or CSW

System Components Targeted for Software Development

* An NGDS Node application, or node-in-a-box, which assists a data provider in sharing their data using the appropriate NGDS standards and protocols, as well as in describing their data using the system’s metadata standards
* A web-application capable of creating and dereferencing URIs for resources in the system
* A single, aggregating catalog that maintains a registry of data-providing nodes in the system, and provides a single point of search for data from the entire system
* A user-centered, entry-point web-application where data consumers go to find, evaluate, explore and acquire data in the NGDS

Additional Requirements Imposed by NGDS Architecture

* Standalone system component implies requirement for a service-based approach to managing URIs that can be utilized by instances of the NGDS Node application
* Generates URIs based on a standard NGDS syntax

Authentication Requirements

* URIs categorized by namespaces, which are likely to correlate closely with instances of NGDS Node applications
* User authentication mechanism with permissions to create/edit tied to namespaces

Functional Requirements

* URIs support regular expression pattern based redirection
* URIs implement HTTP content-negotiation for retrieval of different representations of the same resource

# System Architecture

The framework for implementing data handling requirements is a community of data providers exposing information through standardized internet-accessible interfaces (services), a community of software developers building applications that will utilize the information resources available to the community, and a community of users taking advantage of the software and information to develop geothermal resources. The service inventory would be focused on entity services that provide information resources. As used here, an entity service is a service that provides a requested resource packaged in some interchange format in response to a request, as opposed to a functional service that takes some input package of information and produces an output response according to some processing logic operating on the input information. A key component is the catalog service—an entity service through which data providers register the availability of resources, and users discover, evaluate, and access resources. The system architecture will be described in terms of the functional components shown in Figure 1. These are discussed in the following sections.

## Functional components

Figure . Functional components of National Geothermal Data System. A variety of implementation choices are available for each of the components. Components on the left are mostly hosted by system servers, and interact with the client components on the right through a collection of interfaces defined by the service profiles.

### Catalog

A NGDS catalog component implements one or more protocols for searching a metadata store and returning metadata. At least one of the implemented protocols and interchange formats used for delivering metadata must conform to an NGDS specification. Initial catalog testing and prototypes are using the Open Geospatial Consortium Catalog Service for the Web (CSW), but other protocols such as the Open Archive Initiative Protocol for Metadata Harvest (OAI-PMH) or the OpenSearch protocol may also prove to be useful. The CSW was selected for initial development work because it operates in the same framework as the other Open Geospatial Consortium services being tested for data delivery (the Web Map Service and Web Feature Service), is designed for geospatial data, and has a variety of free, open-source software projects developing clients and servers for the protocol, as well as a variety of commercial products (including ESRI ArcGIS) that are implementing the protocol. The CSW service requires all conformant implementations to return metadata using a simple XML encoding of the Dublin Core Elements and Terms (csw:record), and defines a collection of metadata content elements as core queryable and returnable elements (see OGC 07-006r1). The base CSW specification adds a bounding box as a core queryable requirement for any CSW catalog. Any CSW server must be able to search for criteria based on core queryable elements, and must include the core returnable elements in csw:record XML response documents (although element values may be nil). In addition a CSW service can offer any other xml schema for metadata content, and in the geospatial community, the most widely used profile is for the ISO 19115/19115 metadata. Use of this metadata schema allows richer metadata content that enables greater automation of access to resources. NGDS Catalog instances may be implemented with various software and hardware configurations on any node in the system. To be an NGDS compatible/compliant catalog, the only requirement is that they implement an NGDS catalog service profile, and provide metadata in at least one output format schema and profile that conforms to an NGDS metadata interchange specification.

A metadata content requirement recommendation is being considered by the NGDS separately, and when a metadata content model is adopted for NGDS, encoding profiles for csw:record will need to be established. The USGIN project has developed a USGIN ISO metadata profile for encoding of those recommendations using ISO 19139 encoding of the ISO19115/119 metadata content model. This scheme includes additional metadata attributes and elements for more in depth metadata. Encoding of metadata using the ATOM publishing protocol (<http://tools.ietf.org/html/rfc5023>) has recently been utilized extensively for describing network resources using a simplified scheme similar to csw:record, but with more structured XML to promote greater interoperability. The intention of the USGIN approach is that a small number of these encoding schemes would be adopted, with mappings allowing lossless conversion of content between schemes, allowing implementation of software metadata clients with advanced functionality to streamline user access to the actual described resources.

### Document repository

Data in documents will be accessed via URL from document repositories, which are basically web-accessible file systems. In this context, ‘document’ is used in a very general way as a packaged body of intellectual work with an author (or editor, compiler, or similar originating role), a title, and some status with respect to Review/authority/quality. Documents can be packaged in a single file (e.g. a MS Word document) or a group of related, linked digital files (e.g. ESRI Shape file). Documents provide a straightforward path to get data online quickly and easily for the data provider, but if this approach is used for datasets (e.g. Excel spreadsheets, Microsoft Access databases), it requires the data consumer to do all data integration work themselves. In addition, for the datasets to be useful for data consumers, the metadata descriptions must clearly define the entities and attributes (or features and properties) of the datasets such that users can understand their meaning.

Many options are available for implementing document repositories, including DSpace (FOSS, http://www.dspace.org/), OCLC ContentDM (commercial), Fedora (http://fedora-commons.org/), and the Drupal-based document repository developed in collaboration with the USGIN project (http://repository.usgin.org). In order to integrate holdings in system document repositories, a system repository must make available metadata for contained resources using a NGDS metadata interchange format that can be inserted into the NGDS catalog system. This metadata must contain the required minimum content to allow discovery and access to any document in an NGDS repository, including a URL that will retrieve the resource.

### Data Servers

A Data Server is any component that implements a service providing data using at least one protocol and interchange format conforming to an NGDS specification. Data service delivery of content differs from the simpler document-based delivery because it requires that the format and content delivered will conform to some know set of rules, allowing software to interact directly with the data server to facilitate user acquisition and integration of data into their work environment.

Data delivery through a service requires the service provider to perform any necessary data integration operations to get content into the schema conforming to the service profile. This requires more work for the data provider than the simpler document deliver approach, and thus will have to be implemented incrementally based on the quantity and significance of various data items. Data types that are deemed suitable for service delivery will have NGDS protocols, interchange formats, and vocabularies defined to enable automated access to those data.

Since many of the data types are associated with geographically located features, the Open Geospatial Consortium Web Feature Service (WFS) is proposed as the starting point for implementation of feature services. This protocol uses GML geometry for location description, and allows feature types to be defined that are characterized by feature specific xml schema.

A number of international efforts are under way to develop specifications for data interchange of geoscience information (GeoSciML), and basic observation and measurement data (ISO19156). These xml schema are very flexible to allow representation of a wide range of content, but are thus correspondingly complex. Currently there are no client applications that can do more that transform complex xml to html for display.

Thus, services are defined using simple xml schema with string and numeric-valued elements. These services can be consumed by existing clients like ArcMap and Quantum GIS. Simple feature schema will be compatible with GeoSciML, ISO specifications, and other complex standard schema to the degree that is practical. As clients are developed for richer-content complex feature services, the NGDS can adopt more complex, information-rich schema. There are also a number of other data formats in use in related communities for geoscience information interchange, including WaterML in use by the CUAHSI project, NetCDF, which is widely used for large numeric data sets in the atmospheric and remote sensing communities, and an xml markup developed for geochemical data by the EarthChem project. NGDS has used such schemas as a basis for constructing current information exchanges so that uniformity and interoperability in the science community are more likely to be achieved.

### Database and File System

Various databases and file systems accessed by server applications will house the actual system resources. For security and simplicity, these will probably not be directly accessible for system users, but will be accessed through NGDS service interfaces. Many user applications may also have local data stores, in databases or file systems, used to cache resources obtained from the system for offline usage, better performance, and reliability.

### Clients

The client applications implement most of the desktop analytical and search functionality required by the system. The user interface developed in CKAN (see section below) ingests OGC CSW endpoints like that of the state geothermal catalog for resource discovery and evaluation, and utilize NGDS services and repositories for data access. The system is built to encourage client-side development as discussed in *Data discovery* section above.

## System deployment

### Nodes

Any server that is internet accessible and implements one or more NGDS services, including document repositories containing files indexed by metadata in NGDS catalogs, is effectively a node in the system (Figure 2). Each node will implement one or more of the abstract components shown in Figure 1, and will need to register public resources available at that node in the catalog system.

The deployment diagram indicates a key aspect of the system—the user client software interacts with components on the server side through a pipe labeled “NGDS services.” This connection represents any and all service protocols used to link clients and data servers in the system. These services define interfaces that decouple the clients and servers. Upgrades or modifications to client or server software that do not change the operations and behavior of their service interface will not break the system. This loose coupling is a key design feature necessary to allow the system to evolve as technology and user requirements change.

Figure 2. Deployment of components to nodes in the system. Core nodes will implement special functions, including archives, system specification repositories, and registries of identifiers, as well as standard catalog and data services. Other nodes will implement catalog and data services, and may provide applications that utilize data resources as well. Some applications may provide tightly coupled (client and server specific) linkages to data stores, but these are considered interim solutions because they violate the open access philosophy of the system.

Figure Deployment

Figure Deployment of System components.

### NIAB (Node-In-A-Box) Software Stack for Nodes and Central Aggregator

NGDS developed only one software stack that provides the functionality for both central node and Node-in-a-Box. The reason for this decision is that so many features of both node types are the same that it is easier to develop one software stack and configure the behavior of the system with a configuration file. In order to do so a new parameter to CKAN’s configuration file (“development.ini”) that defines the behavior of a node during startup was added. The node can either be configured as Central Node or Node- in-a-Box. As a Central Node the NGDS software provides the harvesting but no uploading capability while as a Node-in-a-Box the NGDS software provides content uploading but no harvesting capability.

### Software Operating System Support

Ubuntu 12.04 LTS was chosen as the NGDS reference platform. Due to the nature of NGDS, most users will most likely want to run it on a Linux Operating System. Ubuntu is a well-known and well-documented Linux OS. Also, CKAN is optimized for Ubuntu or other Debian-based Linux distributions.

NGDS has been developed with Ubuntu Version 12.04 LTS because it has a long support cycle. When the next LTS version becomes available the NGDS development environment will be upgraded. Installation files, etc. are written in such a way that they will likely work on any Debian-based Linux version. Therefore, porting to other Linux platforms is a minor effort. NGDS is also frequently tested and installed on the Mac OS X platform but NGDS is not test installed the system on Windows. Most likely it is possible to start the system on Windows but to fully support that would drain too many development resources and it is questionable if this form of testing adds much value.

### Software Base Back-end Technologies

NGDS relies heavily on Python and CKAN. This defines the architecture of NGDS to a great extend because CKAN is a framework with well-defined extension points where NGDS functionality has been added. Further, CKAN is used as-is without modifying the CKAN core. Bugs are found in the CKAN core are reported and tracked through the Open Knowledge Foundation’s support team and their tools.

Since CKAN is targeted towards Postgres, NGDS’s reference database is also Postgres. The PostGIS extension of Postgres is used for geographic features.

In order to serve OGC services NGDS uses Geoserver which runs in its default setting on jetty.

For indexing of Metadata and full-text indexing NGDS uses SOLR (on jetty). SOLR is configured according to the CKAN recommendations.

### Software Base Front-end Technologies

NGDS uses HTML5 and CSS3 as the base technology for the frontend. HTML5 and CSS3 are now supported by all major browsers (even later versions of Internet Explorer). In addition NGDS uses various JavaScript Libraries. Specifically, NGDS uses JQuery and various libraries based on JQuery. Furthermore, NGDS uses CSSless in order to reduce CSS complexity. At the time of writing of this document these are the state-of-the-art technologies for developing Web applications.

For the production of the HTML pages, NGDS uses the Jinja2 templating system that is built into CKAN. The templating system can be compared to PHP and allows to bring backend-information (made available via Python) into the frontend HTML content.

Jinja2 is very flexible and works well in the CKAN environment. The reason for using Jinja2 is that it is the reference solution for CKAN. Building other templating systems into CKAN would be extra effort and causes problems regarding maintainability.

In many cases it is possible to create the page by either using Jinja2 templating or JavaScript. Jinja2 is preferred rather than JavaScript whenever possible because Python code and Jinja2 code is in general easier to maintain than JavaScript code.

The Leaflet Map Widget is used to present maps ([http://leafletjs.com/).](http://leafletjs.com/)) This widget is currently popular and has an active developer community. OpenLayers (<http://openlayers.org/)> was considered but due to higher experience with Leaflet among the NGDS developers, it was decided to go with Leaflet.

Naturally it is not possible to simply replace Leaflet with another map widget solution. However to the extent possible, the Leaflet-specific code is encapsulated in separate classes and modules so that the classes can be replaced in the future in case that the map widget needs to be exchanged.

# Data Acquisition Plan

This data acquisition plan is a road map for bringing data into the information infrastructure that is the foundation of the NGDS. The intention is to get the NGDS off the ground, with useful data content, as quickly as possible by using existing, tested Open Geospatial Consortium services, particularly Web Map Service (WMS) and Web Feature Service (WFS). In a nutshell, the steps in this plan are:

1. Identify the kinds of information to be made available through the system.
2. Prioritize acquisition according to availability, importance for geothermal resource evaluation and development, and difficulty of acquisition.
3. Make data resources accessible
   1. For document based resources and datasets that do not have specifications for interchange protocols, data schema, and file format: create metadata for resource and make resource available in a web-accessible location linked to from the metadata.
   2. For high value datasets with sufficient volume, design and implement xml schema based on any applicable standards to use as an interchange format in WFS service response documents, and make the data available through WFS service. Metadata describing service function and content go in catalog.
   3. Map-based portrayals of information can be made available as documents, and as WMS service layers. Metadata describing map content and distribution points go in catalog.

Data types for which NGDS data acquisition services and interchange formats aremade available in user-defined data files described by metadata in the system catalog and placed in web-accessible servers. Standardization of automated, interoperable data acquisition via services and community interchange formats have been developed for NGDS as an ongoing, living process for the geoscience community. The infrastructure that supports the normative schema locations is at <https://github.com/usgin/modelmanager> which houses the code base that supports the Django NGDS schemas management site <http://schemas.usgin.org/models/>. The information exchanges are developed and maintained at <https://github.com/usgin-models>, where GitHub repositories exist for each given data type. These tagged versions of schemas are then implemented at the aforementioned repositories and sites for use in the system. The community of users is tasked with continuing development and versioning of new or existing information exchanges as needed. A detailed description of the workflow for defining a new information exchange can be found at <https://github.com/usgin/usginspecs/wiki/Define-New-Information-Exchange>, with a detailed description of NGDS standards for constructing them at https://github.com/usgin/usginspecs/wiki/Content-Model-Guidelines.These documents allow for interoperable data to be presented to the system using standardized protocols, interchange formats, and vocabularies, while allowing the user community (data providers and consumers) to continue developing information exchanges with standardized units of measure and controlled vocabularies. Interoperability means in practice that software will use the same access protocol for a given kind of information from any NGDS data provider, without any provider-specific customization. Some important requirements include:

1. Ensure interoperability among data sets with members adopting common standards and protocols.
2. The data schema must be vetted with stakeholders
3. Data schema for interchange formats, and instance documents based on these schema must be versioned, such that expanded or modified versions can be introduced without disrupting working systems.

The process of identifying kinds of information to be made available will be pursued on two fronts. NGDS consortium members were polled in January and February, 2010 to get an inventory of the resources that they will be contributing to the system, but the results were limited in terms of specifics, mostly recognizing scanned well logs and other kinds of documents. The data resource inventory has been continuing through verbal interviews by AZGS development staff with information managers at several of the organizations. With the initiation of the AASG geothermal data project, state geological surveys were polled yielding a larger body of data resources to be made available through the system. The evolution of this inventory has continued as states developed plans for data contributions and pulled survey legacy data, then submitted those resources to the project. Input from the SMU/Siemens Geothermal Data compilation project was also factored in, yeilding the currect collection of data types in information exchanges. A current list of those information exchanges can be found and downloaded from http://schemas.usgin.org/models/.

The data acquisition process will be planned to focus on delivering information to enable use cases being implemented by the Geothermal Desktop application in order to make utilization of implemented functionality immediately useful.

## File based data

File-based data access will be the option of choice for text documents, but will also be used for data sets that do not have a standard interchange protocol and file formats defined. Some tabular file formats may already be in use, or be specified by groups of users to simplify exchange of some kinds of information, and if widely used these would be obvious candidates for system interchange formats. The recommended metadata for file-based (document) resources is designed to allow discovery, evaluation of the resource based on text description, and access to the resource via a web link (URL).

### Data to be scanned

Reports, logs, maps and other documents pertinent to geothermal energy exploration, evaluation, development, and production that exist in hard copy but are not available online may be converted to digital form by scanning to create digital image files. If the resource is a map, it should be georeferenced (geoTiff or world file) if possible. Preferred document formats are pdf, tif, jpg, or png. File formats that are specific to particular (especially proprietary) software are undesirable and their use will need to be justified and approved by the project management. OCR processing of text to make Adobe Acrobat files searchable is highly desirable. Georeferenced map images ideally will be published through a Web Map Service (WMS) as well as accessed from document repositories. Deliverable digital documents must be publicly available online, and registered in the NGDS metadata catalog. A prototype document repository, implemented using Drupal software is available for deployment by data providers that do not currently have such an online repository (<http://repository.usgin.org/>). This application also supports production of metadata meeting NGDS requirements. Instructions for deployment are available at <http://lab.usgin.org/groups/drupal-development/creating-document-repository-drupal>.

## Online Digital data

Implementation of online data services will involve several steps. First, an application profile for the service or services to deliver a particular kind of data will have to be developed. In most cases, we anticipate that existing standard services like the Open Geospatial Consortium (OGC) Web Feature, Map, or Coverage services (WFS, WMS, WCS) will provide the necessary protocol for services we require. NGDS technical teams will need to develop content models specifying the details of how a particular feature or observation will be described, and an encoding scheme to serialize this information for transmission over the web. Once a profile is in place for a particular data resource, the next step is working with the data providing organizations to map existing data to the interchange format and to implement a service to expose the data.

The actual mechanics of bringing particular datasets online will be dependent of the format of existing data, and the IT resources of the data owner. Some organizations may choose to implement web services on their own servers to expose datasets, others may choose to work with a partner that has better IT support to host services.

The second part of the online service implementation and deployment is registering the new data service with the catalog system. This will require creating a metadata record for the service, and loading it into a catalog server or web-accessible directory that is harvested by the NGDS catalog system, such that the fact of the service’s existence, and information to evaluate and access the service becomes available to the community. The data acquisition process will thus need to include guidance on what kind of metadata will be required to register resources with the catalog system to make them available.

For online data services, registration of a dataset in the catalog, and its availability online will constitute ‘data acquisition’. Thus, implementation of the catalog as an operational service will need to be one of the first steps in system implementation. AZGS has developed a prototype catalog, implementing the CSW 2.0.2 catalog service using the ESRI GeoPortal, as well as tools for individual metadata record creation using Drupal (<http://mw.usgin.org>) and for bulk loading using a MS Excel Workbook (http://www.stategeothermaldata.org/data\_delivery/content\_models/metadata\_template).

# Technical discussion

## Data delivery options

Participants have two options on how to make their data available:

* Register files in an NGDS-compliant document repository; submit metadata to NGDS-compliant catalog. If the files contain datasets, then the structure of the data (entities, attributes, vocabulary) should be described in the metadata such that someone using the file dataset can figure out what they’ve got.
* Implement a web service for direct online access to the data. Submit metadata to NGDS-compliant catalog.

Data will be considered part of the NGDS when it is locatable using the NGDS core catalog, and accessible via the web according to procedures described in the metadata record obtained from the NGDS core catalog.

## Metadata

*Metadata* should be created and submitted for any resource that is meant to be accessible individually via the web.

*Individual documents* require one metadata record per document. Some document types may consist of a bundle of files, e.g. ESRI shape file. In general these should be bundled into a single file like a zip archive or UNIX tar file. The metadata must include the URL at which the document can be accessed. These documents might be scans of well logs, scanned reports or publications, or data in a spreadsheet, such as an Excel file.

*Datasets* include internal record level source information, documenting details of observation or measurement procedure and other information specific to a particular data type. This includes information such as location, data and time of observations, and the source of the data. These metadata are delivered with the data, and only summarized in the dataset metadata that are published to the NGDS-compliant catalog.

The required metadata content will be documented in a metadata specification document that has been submitted for Technical Working Group comment and review.

# 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 protocol (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 mainstay of most interaction in the NGDS.

The OpenGeospatial Consortium Catalog Service for the Web (CSW), currently at version 2.0.2 is proposed for catalog search and discovery service. The lowest common denominator metadata interchange format using this service is an encoding of the Dublin Core elements and Dublin Core text extensions (schema at <http://schemas.opengis.net/csw/2.0.2/rec-dcmes.xsd>, <http://schemas.opengis.net/csw/2.0.2/rec-dcterms.xsd>), and the NGDS needs to adopt a best practice recommendation for using this metadata encoding to achieve interoperability between metadata provided by various servers. For more in-depth metadata, use of the USGIN profile for ISO metadata is proposed. All CSW implementations we are familiar with implement the CSW ISO profile, and various groups (NOAA, Univ. of Zaragoza Spain) have worked out software to translate FGDC CSDGM to ISO 19139 (although the process is not perfect).

Initial data services can be implemented using WFS 1.1.1 simple feature services, selecting a few widely available and geothermally interesting datasets. Based on data compilations thus far, the AASG Geothermal data project has implemented borehole temperature observation services, Quaternary fault feature services, water chemistry observation services and volcanic vent feature services (see http://services.azgs.az.gov/ArcGIS/rest/services/aasggeothermal). The content model and xml schema used for data interchange in these services will need to be reviewed by the full NGDS.

# Glossary

Definitions here are meant to clarify the usage of terms in this document.

**Artifact**: A thing created by humans, usually for some practical purpose. (Source: http://www.merriam-webster.com/dictionary/artifact)

**Attribute**: A binding between a property, a data type, and a data item; an implementation of a property.

**Cardinality**: A constraint on the number of instances of assigned property values associated with an individual data item. A cardinality of 1 indicates exactly one value is required; 0..1 indicates an optional single value; 1..n indicates that one or more values is required; 0..n indicates that a value is optional, and multiple values may be specified.

**Content model**: A model that identifies and defines the data items and the properties (with cardinality) associated with each data item.

**Data integration**: the process matching field or element names in the schema for various data sets, selecting those that contain the information of interest, and merging content into a single data set with consistent usage of vocabulary and units of measure in a standardized collection of fields or elements.

**Data item**: An identifiable unit of information. Generally represents some entity in the world.

**Data type**: A specification of the representation of a single value in an information system, using integer, floating point, string, Boolean.

**Entity service**: a service that provides a requested resource packaged in some interchange format in response to a request, as opposed to a

**Feature type**: Type for representing a feature.

**Feature**: An information resource representing some identifiable thing of interest in the world.

**Functional service**: a service that takes some input package of information (message) and produces an output response (message) according to some processing logic operating on the input information.

**Information resource**: A resource that can be transmitted electronically.

**Interface**: a point of interaction between components, typically defined by a protocol for transmitting messages and a collection of method names and parameter specifications used to invoke operations executed by a component.

**Interoperability**: "The capability to communicate, execute programs, or transfer data among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units." ISO/IEC 2382-01 (SC36 Secretariat, 2003)

**Observation:** an information resource representing the event of observing or measuring and recording properties of some feature (Open Geospatial Consortium, Observations and Measurements (O&M), <http://www.opengeospatial.org/standards/om>). Observations represent the basic data that are the foundation for scientific knowledge.

**Operation:** an individual process that a software component may execute

**Property**: A phenomenon that is inherent in the nature of some other phenomenon, and may be used to characterize it by specifying a value.

**Protocol**: A set of rules which is used by computers to communicate with each other across a network (<http://en.wikipedia.org/wiki/Network_protocol>).

**Representation**: A binding between a symbol or collection of symbols (in language, text, graphics, computer bits, etc.) and a human concept or resource.

**Resource**: An identifiable thing that fulfills a requirement. Usage here is close to definition used in RDF (<http://www.w3.org/TR/REC-rdf-syntax>), generalized from ISO19115, which defines resource as an ‘asset or means that fulfills a requirement’ without defining asset or means. "An object or artifact that is described by a record in the information model of a catalogue." (OGC 07-006r1)

**Schema**: A formally structured representation of a conceptualization. A model presented using some specific notation.

**Service**: A system that provides one or more functions via a network interface designed for machine interaction; utilization involves some agent making a request and possibly providing some input, at which point the service executes the requested procedure with some predictable result

**Specification**: A document that describes the technical characteristics of an artifact, possibly including a description of what it should do, or an explicit set of requirements that it must satisfy. (Based on <http://en.wikipedia.org/wiki/Specification>).

**Type**: Specification of a collection of attributes and cardinalities for those attributes used to represent a data item.

# References

Dublin Core Initiative, 2008-01-14 Dublin core Metadata Element Set, Version 1.1: Dublin Core Metadata Initiative, accessed at http://dublincore.org/documents/dces/.

Franklin, Michael, Halevy, Alon, and Maier, David, 2005, From databases to dataspaces: a new ab-straction for information management: ACM SIGMOD Record, V. 34, No. 4, ISSN:0163-5808.

Open Geospatial Consortium Catalog Service for the Web (CSW): http://www.opengeospatial.org/standards/cat

Open Archive Initiative Protocol for Metadata Harvest (OAI-PMH): http://www.openarchives.org/pmh/

OpenSearch protocol: http://www.opensearch.org/Home

ISO 19115:2005, Geographic information - Metadata

ISO 19115/Cor.1:2006, Geographic information – Metadata, Technical Corrigendum

ISO 19119:2005, Geographic information - Services

ISO 19119:2005/Amd 1:2008, Extensions of the service metadata model ISO 19108 designates:

OGC 07-006r1¸ OpenGIS Catalog Services Specification version 2.0.2, Corrigendum 2 release, 2007

OGC 07-045, OpenGIS Catalogue Services Specification 2.0.2 - ISO Metadata Application Profile, Version 1.0.0, 2007

USGIN Team, 2010, Use of ISO 19139 xml schema to describe geoscience information resources v1.1, Current version at http://repository.usgin.org/uri\_gin/usgin/dlio/337.

USGIN Team, 2010, Metadata Recommendations for Geoscience Resources, current version available at http://repository.usgin.org/uri\_gin/usgin/dlio/335.