PROCEEDINGS, Thirty-Eighth Workshop on Geothermal Reservoir Engineering
Stanford University, Stanford, California, February 11-13, 2013
SGP-TR-198

NATIONAL GEOTHERMAL DATA SYSTEM (NGDS) GEOTHERMAL DATA DOMAIN: ASSESSMENT OF GEOTHERMAL COMMUNITY DATA NEEDS

Arlene Anderson[1], David Blackwell[2], Cathy Chickering[2,] Toni Boyd[3], Roland Horne[4], Matthew MacKenzie[5], Joseph Moore[6], Duane Nickull[5], Stephen Richard[7], Lisa A.Shevenell[8]

[1] United States Department of Energy
e-mail: arlene.anderson@ee.doe.gov

[2] Southern Methodist University
e-mail: blackwell@mail.smu.edu
e-mail: catherine@mail.smu.edu

[3] Oregon Institute of Technology Geo-Heat Center
e-mail: toni.boyd@oit.edu

[4] Stanford University
e-mail: horne@stanford.edu

[5] Uberity Technology Corporation
e-mail: matt@uberity.com, duane@uberity.com

[6] University of Utah, Energy & Geoscience Institute
e-mail: jmoore@egi.utah.edu

[7] Arizona Geological Survey,
e-mail: steve.richard@azgs.az.gov

[8] University of Nevada, Reno
e-mail: lisaas@unr.edu

ABSTRACT

To satisfy the critical need for geothermal data to advance geothermal energy as a viable renewable energy contender, the U.S. Department of Energy is investing in the development of the National Geothermal Data System (NGDS). This paper outlines efforts among geothermal data providers nationwide to supply cutting edge geo-informatics. NGDS geothermal data acquisition, delivery, and methodology are discussed. In particular, this paper addresses the various types of data required to effectively assess geothermal energy potential and why simple links to existing data are insufficient. To create a platform for ready access by all geothermal stakeholders, the NGDS includes a work plan that addresses data assets and resources of interest to users, a survey of data providers, data content models, and how data will be exchanged and promoted, as well as lessons learned within the geothermal community.

INTRODUCTION

Geothermal energy, or literally, the thermal energy of the earth, is often used as a term to refer to conversion of the earth's thermal energy into electricity. Present geothermal power generation comes from high-temperature hydrothermal systems, the 'low-hanging fruit' of geothermal electrical potential. Newer techniques, including Enhanced or 'Engineered' Geothermal Systems (EGS) offer the opportunity to extend use of geothermal resources to larger areas of the western U.S., as well as new geographic areas of the U.S.

Through the American Recovery and Reinvestment Act of 2009, the U.S. Department of Energy (DOE) funded the development of a National Geothermal Data System. In early 2008 DOE issued a funding opportunity announcement to develop a "National Geothermal Database" to overcome barriers to the development of geothermal energy facilities and enable additional investment in conventional and Enhanced Geothermal Systems (EGS). Based on the proposals received and subsequent technical review,

DOE funded an effort to create a web-based National Geothermal Data System for all publically accessible geothermal data. Data needs span all geothermal resources and applications including geothermal electricity production as well as direct use applications. Geothermal data is being contributed by industry, academic and national laboratory researchers, and by state and federal agencies. While the focus is on domestic data critical to identifying geothermal potential and characterizing geothermal reservoirs, international data sources may be included especially where such data and information can be utilized or benchmarked to help develop domestic geothermal resources. The system is being implemented using a federated, Service Oriented Architecture (SOA) based on the U.S. Geosciences Information Network (USGIN) (http://usgin.org/). The DOE adopted the US Government Accountability Office (GAO) (http://www.gao.gov/new. items/d06629.pdf) best practices for software development featuring an agile development process that incorporates the latest informatics technology and standards into the system design. NGDS has also adopted the International Organization for Standardization (ISO) (http://www.iso.org) metadata standards for the system catalog.

Data analysis for geothermal resource development presents a highly complex challenge where: "The rate-limiting step for all geothermal development is proving the resource – i.e., having sufficient geoscientific and exploration drilling data to be certain of a certain level of output" (Bloomberg New Energy Finance, 2012).

A variety of data is required to ascertain whether a potential geothermal energy site should be developed for production: composition and hydrologic properties of materials hosting the thermal energy, proximity to existing power grids, and quantity of thermal energy flowing from the interior of the earth are all primary considerations.

The NGDS will provide critical geothermal-related data that can be easily accessed to:

- Help companies be more (cost and time) effective in exploration, development, and usage of geothermal energy.
- Support a knowledge repository and archive for geothermal data, lessons learned, reports.
- Advance earth sciences by identifying gaps in our knowledge and informing new geographic areas of the U.S.
- Provide a reliable base load energy source of knowledge.
- Increase public awareness of geothermal energy.

These goals can only be accomplished if NGDS provides a quality user experience, and is widely adopted by users in the geothermal community.

There are three targeted user communities for NGDS, and each user group has different goals, needs, and tasks when interacting with NGDS.

- **Data providers** expose information to NGDS through standardized, internet-accessible interfaces and standardized formats.
- **End users or data consumers** utilize NGDS to access data to support their work in geothermal energy exploration and development.
- **Application developers** build applications that utilize the data in NGDS, and make it easier for end-users to interact with the system.

An additional NGDS goal is helping users to understand where geothermal investment will have the best opportunity for success. Through NGDS, users will gain access to tools that can improve the usefulness of geothermal data and information.

Providing simple links to geothermal datasources across the country would only improve knowledge enablement to a limited extent. A non-exhaustive list of reasons why making simple links to existing data sources is inadequate includes:

- Data is in multiple formats, layouts, units, paper versions and not searchable via one central index;
- Some database persisted data is difficult to access, visualize and/or interpret, especially for the business/industry user;
- There is no current ability to link some data to additional geological information or datasets;
- There are inconsistent standards for quality assurance or reliability of data.

In order to structure the records so the data can be linked and interoperable, data and metadata content models and interchange formats were created. To date, twenty-eight geothermal data models have been developed, reviewed and adopted. The Geothermal Domain Committee provided expert input on a prioritized list of data models including geothermal drilling and well log data, aqueous chemistry, geophysical data and active fault maps. Expert input on whether data models include the correct information is critical.

NGDS Data Architecture

The NGDS is not a single database. Rather, it is a unified data access system based on the registration of resources in a shared catalog system using standardized metadata. NGDS has a tiered data access scheme accommodating file-based, non structured, and standards-based structured data delivered using standardized web services and interchange formats. A data resource becomes part of the NGDS system when standard NGDS metadata is created, validated,

and made discoverable through the NGDS catalog system, and the data resource is accessible via procedures specified in the metadata. Much of the information that is or will be registered in the NGDS is unstructured data. Other resources, such as drill cores, may not be available in electronic format. Especially in such cases, metadata is essential to allow NGDS users to be aware that the resources exist.

Figure 1: An example of a resource (drill cores) requiring metadata records indexable via NGDS (Photo courtesy of Energy & Geoscience Institute).

Additional information about the system design is discussed by Clark et al. (2013).

GEOTHERMAL DATA PROVIDERS

The NGDS provides access to information resources on geothermal energy from a national network of data sources (http://geothermaldata.org/). As of the date of this paper, four project teams are collaborating and leveraging efforts that will culminate in the NGDS launch. This includes the NGDS Design & Testing Project; Heat Flow Data Aggregation; State (geological survey) Contributions to NGDS; and the DOE Geothermal Data Repository. Additional information about the GDR is provided by Weers and Anderson (2013).

Under the leadership of Boise State University (BSU), the NGDS Design & Testing Project includes four key geothermal data providers including: the University of Utah Energy & Geosciences Institute; the University of Nevada, Reno; the Stanford Reservoir Engineering Department, and the Oregon Institute of Technology Geo-Heat Center. The Arizona Geological Survey and Siemen's Corporate Research lead the NGDS informatics development.

Two additional NGDS projects focused on data content are: Arizona Geological Survey (AZGS) "State Contributions to NGDS" including data from all 50 state geological surveys and the "Heat Flow Data Aggregation for NGDS Data Development, Collection and Maintenance" project led by the Southern Methodist University's (SMU) Geothermal Laboratory. The SMU consortium includes: the Bureau of Economic Geology (BEG), University of Texas at Austin; Cornell Energy Institute, Cornell University; the Geothermal Resources Council (GRC); MLKay Technologies; Texas Tech University (TTU); the University of North Dakota (UND), and Siemens Corporate Research (SCR).

The SMU node is focused on improving access to information to allow for new interpretation of data, thereby increasing its usefulness for commercial geothermal energy development. The DOE-GDR is another vital node on the NGDS. The GDR is hosted on the Open Energy Information (OpenEI) Platform.

The U.S. Geological Survey (USGS) and other federal agencies that produce geothermal data are potential NGDS data providers. An interagency agreement with the USGS is designed to provide geothermal resource assessment and classification data. As part of the national geothermal resource assessment, USGS has conducted a comprehensive survey of the available information on geothermal systems and an extensive set of geothermal databases. These databases include chemical analyses of water and gas samples, heat flow measurements, gravity and magnetic surveys, geologic maps, seismicity catalogues, seismic surveys, drilling records and other relevant exploration and development data, including consultants' reports and interpretive studies.

USGS personnel have combined many of these supporting geological, geophysical, geochemical, and

hydrologic datasets into Geographic Information Systems (GIS) databases and maps for analysis and publication. The current USGS data provision strategy is to provide online access to data and reports either through direct delivery of data to the DOE GDR or through state geological survey NGDS nodes. The USGS will also provide the information necessary to ensure that geothermal resource data will be current and available. Examples of the data include geothermal resource assessment and derivative products, such as GIS maps, low-temperature data series and related publications. As new data is acquired and geospatial products and data are developed, they will be sent through USGS for review and subsequently provided to the NGDS.

Table 1 summarizes the work plan for deliverable data assets from NGDS Design & Testing Project partners. Abundant data has also been incorporated into the NGDS by the AASG "State Contributions to NGDS" project. The work plans and progress can be monitored at http://www.stategeothermal- data.org/progress/aasg_tracking_map. A list of data delivered by that project is available at http://repository.stategeothermal-data.org/ repository/browse/. An interim search interface for the NGDS is accessible at http://search.geothermaldata.org/.

THE NGDS DATA MODEL

The NGDS is based on a model (Figure 2) that uses the top class ***NGDS_Resource*** to denote any resource within the NGDS system NGDS_Resource's can be further classified as ***Data Resources***, ***Metadata*** or ***Annotations***. NGDS Data Resources represent the actual resources of interest to end users. An example of an NGDS Data Resource might be a spreadsheet, using Comma Separated Values (CSV), showing the temperature of a well at different depths or a physical drill core sample. Every NGDS data resource must be described by at least one metadata record, as described below. The onus of data maintenance is shifted towards organizations having responsibility for data management and preservation. By documenting data schema, encoding formats and practices, data can be put into the 'data integration' format 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. As previously referenced, the NGDS will initially use twenty-eight data content models and interchange formats, as well as a standard metadata scheme. An integrated data access portal application is in development.

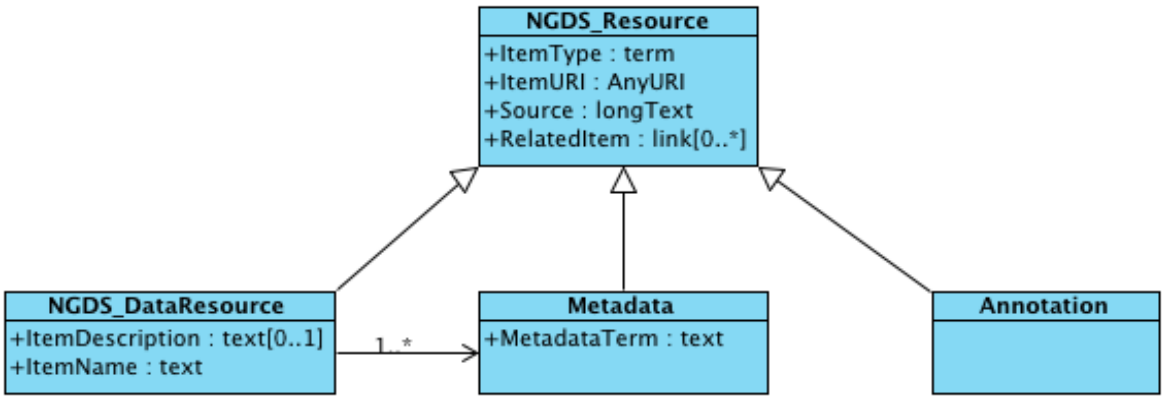

Figure 2: The high level NGDS data model.

Table 1: *Summary of deliverable data items from NGDS development and testing project participants.*

Group	Data Item
Stanford	Bibliographic Database for Proceedings from the annual Stanford Geothermal Workshop count: 2118 metadata records with location keywords
	Metadata Records for 3 Adsorption Data publications
GeoHeat Center, Oregon Institute of Technology (OIT)	717 Technical Papers and bulletin articles online, with NGDS metadata records
	Metadata for 4185 documents in the Geo-Heat Center Library
	Documentation and registration of data set describing 554 Geothermal Wells in Klamath Falls area
	Documentation and registration of data set describing 404 Co-located Sites
	In cooperation with Siemens Corporate Research (SCR) and University of Nevada, Reno (UNR), thermal springs and borehole temperatures will be de-duplicated for the 16 western states, processing non-standard location information, and served in the NGDS content model as the OGC's Web Map Services (WMS, OGC 07-063r1) and Web Feature Services (WFS, OGC 09-025r1 and ISO/DIS 19142).
	Documents and data related to the Klamath Falls #57310 project will be scanned and publicly accessible online with metadata.
	Metadata for GeoHeat software Tools and Spreadsheets.
University of Utah, Energy & Geoscience Institute (EGI)	2635 Scanned well logs indexed in NGDS Well Log Observation Content Model.
	9010 scanned reports, articles, maps, charts and graphs with metadata.
	Geothermal Sample Library samples registered with System for Earth Sample Registration (SESAR - http://www.geosamples.org/), and correlated with well log and well header data sets
	Create metadata for more than 1000 Scanned Documents
	Catalog and scan 20 boxes of well logs.
University of Nevada, Nevada Bureau of Mines and Geology (NBMG)	Metadata *for* more than 400 known publications and grey literature relevant to geothermal exploration and development in Nevada
	More than 2000 documents (notices, permits, gray literature) to be scanned and placed online with metadata records
	Approx. 150 1:24k scale geologic maps to be scanned and geo-referenced, with metadata
	Map and report describing all exploration activity reported in 2012 will be scanned, put online, with metadata
	Metadata for more than 179 existing geologic, geophysical and geochemical data sets relevant to geothermal assessment. Update NBMG Geothermal web map applications to operate with Tier 3 NGDS services.
	NBMG Geothermal map applications will be updated to operate with NGDS services and integrated with NGDS applications being developed by Siemens.

Data Tiers

NGDS was designed to use a tiered data delivery scheme that allows the necessary flexibility to accommodate unmanaged legacy data in whatever form it is available, as well as high value data in standardized content models and/or interchange formats. The system uses a community governance scheme to adopt new interchange formats, and provides a repository where the specifications for each data exchange are available to all.

In order to make the incorporation of a large quantity and variety of data in the NGDS, a tiered data acquisition scheme has been used.

- **Tier 1: Unstructured** — represents file based resources such as unstructured data in text and images, requires a user to extract data for analysis.
- **Tier 2: Structured, but not standardized** — represent data structured in proprietary formats that are not conformant with a standard NGDS content model. Data in this tier would need to be transformed in some fashion by a data consumer in order to integrate with NGDS-standard datasets.
- **Tier 3: Structured, standardized** — data published in the NGDS standardized protocols and interchange formats supported by NGDS content model.

A large part of the available resources are scanned images of legacy reports, maps, and other figures that are registered with metadata and made available as Tier 1 resources. Tier 2 allows registration of existing structured datasets that are not in standard NGDS content models and interchange formats. This is not a preferred data acquisition approach, but is expedient and useful for unique datasets that have only a single instance.

Tier 3 data acquisition is the preferred scheme, but because of the additional effort required to edit and review datasets to get them into the standard interchange format, it has been necessary to prioritize effort. This was done by first surveying the data provider community to determine the types of structured data that they actually have available for inclusion in the NGDS. The team then informally polled geothermal exploration and development practitioners (mostly in the State Geological Survey community) to determine which of these types should be prioritized.

Linking Geothermal Data Providers Through Service Protocols

A protocol is a set of rules used by computers to communicate with each other across a network. Since virtually all of the data types identified for NGDS Tier 3 interchange are geographically located features, the OGC Web Feature Service (WFS) (Vretanos, 2005) is being used as the data service protocol and the OGC's Web Map Service (WMS) (De La Beaujardiere, 2006) is used as a standard protocol for serving geo-referenced map images over the Internet that are generated by a map server using data from a GIS database.

The WFS protocol uses the OGC Geography Markup Language (GML) (Portele, 2007) geometry for location description, and allows feature types to be defined that are expressed by feature-specific eXtensible Markup Language (XML) schemas. Geographic data is also made available for viewing with geographically enabled software as OGC Web Map Services.

Document-based resources use the standard HyperText Transfer Protocol (HTTP) that is the foundation of the World Wide Web. WFS and WMS are implemented on top of HTTP. Some data providers also make files available using the File Transfer Protocol (FTP). These protocols are both widely used within the geothermal community.

CONTENT MODELS

In the NGDS, content models specify the structure and properties associated with an interchange feature, typically including feature-specific metadata allowing documentation of each data item. Content models are specified independent of interchange formats, the latter being a typed expression of the content model. If data cannot be structured using an existing content model, geothermal community members are invited to propose new models.

Development of content models during the first year of the project has been an organic process. The models have evolved rapidly as production scale data compilation has started.

Content Model Inventory

Various approaches have been used to prioritize the kinds of data that will be implemented using Tier 3 services. 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 continues through verbal interviews with information managers at data provider organizations and with geothermal industry practitioners. With the initiation of the AASG "State Contributions to NGDS" project, state geological surveys were polled yielding a larger body of data resources to be made available through the system. The evolution of the Tier 3 information exchange inventory will continue as NGDS participants develop plans for data contributions, and new projects and participants are factored in.

Content models available to date include (see http://geothermaldata.org/page/ngds-content-models for details on the content models):

- Aqueous Chemistry
- Borehole Temperature Observation Feature
- Data Interchange Content Models
- Direct Use Feature
- Drill Stem Test Observations
- Fault Feature

- Fluid Flux Injection and Disposal
- Geologic Contact Feature
- Geologic Unit Feature
- Geothermal Area
- Geothermal Fluid Production
- Geothermal Power Plant
- Heat Flow
- Heat Pump Facility
- Lithology Interval Log Feature
- Metadata
- Physical Sample
- Powell Cummings Geo-thermometry
- Power Plant Production
- Radiogenic Heat Production
- Seismic Event Hypocenter
- Thermal Conductivity
- Thermal/Hot Spring Feature
- Volcanic Vents
- Well Fluid Production
- Well Header
- Well Log Observation

Additional content models under consideration:

- Daily Drilling Report
- Well completion Information
- Well production hardware
- Surface Alteration
- Subsurface Alteration
- Geophysical Survey Results
-

Figure 3 introduces a more in-depth model of the data item content models used for Tier 3 data. The content models are designed based on this pattern, with a distinction between features (Kottman and Reed, 2009) that represent geographically located real-world phenomena, and observations (Cox, 2010) to represent individual measurements of one or more properties of some real-world phenomena. A Feature typically summarizes the results of multiple observations to characterize something like a fault, a geologic unit, a well, a power plant, or a geothermal area. Observations represent the more granular data, 'raw' data like individual temperature measurements, chemical analyses, or heat flow determinations. Observations may have composite results; for instance an individual well log is considered an observation result from a log run event.

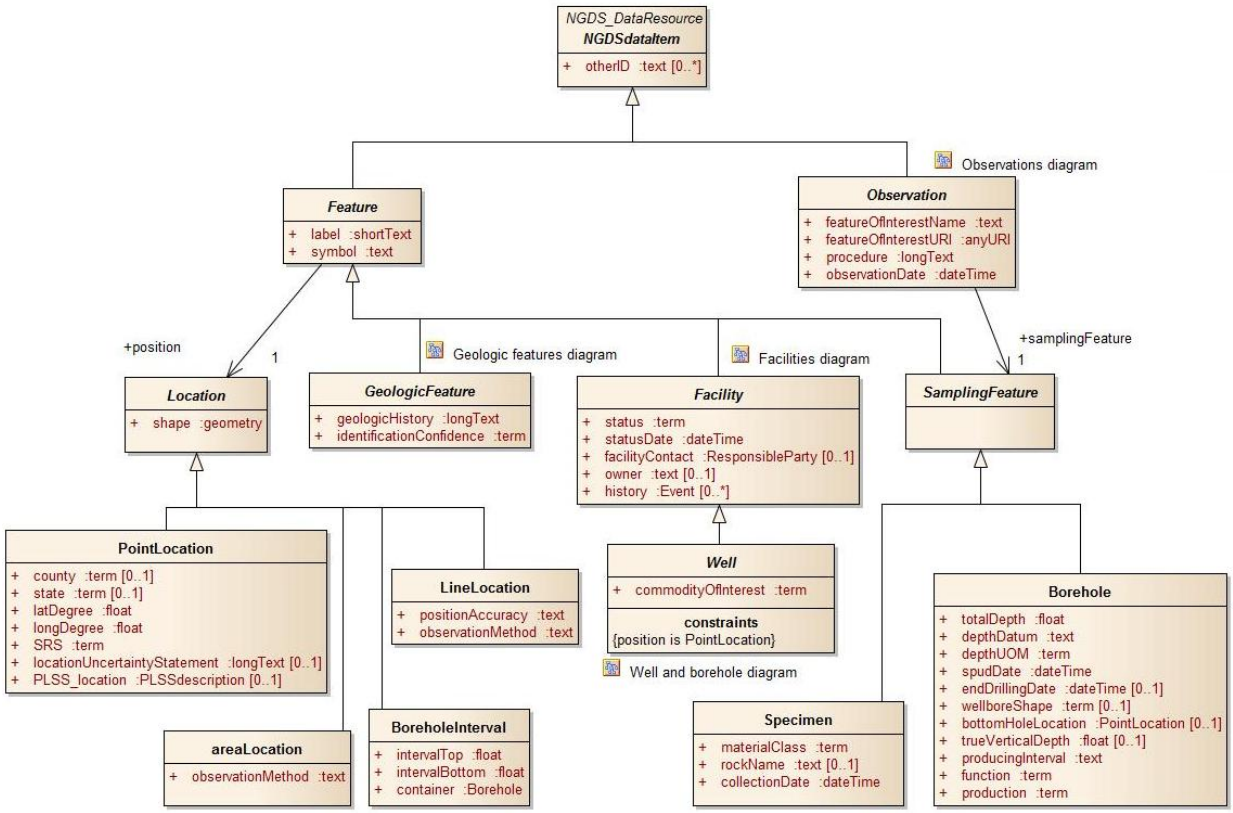

Figure 3: Top level model of NGDS Data Items for Tier 3 data sets.

The key property of a Feature is that it has a spatial-temporal location; only the spatial aspects of location are modeled in Figure 3. Features are subtyped into two broad categories of interest to the geothermal domain. Geologic features are used to represent natural features within the Earth; subtypes include geologic unit outcrop, fault, Quaternary fault, geologic contact, volcanic vent, thermal spring, and geothermal area. Facility is used to represent feature of human origin; subtypes that are currently implemented include well, geothermal direct use site, heat pump facility, and power plant facility.

Another sub-category of Feature, SamplingFeature is used to represent the artifacts that are the immediate target of observations, and serve to geo-locate and contextualize an observation result; specimen (core, rock sample) and boreholes are the dominant sampling features of interest in the NGDS.

Observations represent observed or measured property values that characterize a feature of interest (e.g. a rock unit or geothermal reservoir), have a measurement procedure, are associated with a sampling feature, and have one or more result values. Observation types currently implemented include fluid flux, seismic event, drill stem test, rock chemistry, aqueous chemistry, heat flow, borehole temperature, and borehole lithology interval.

Interchange Formats

In order for information to be exchanged, a content model must be serialized in a form that can be transmitted over a computer network and interpreted by software applications. Interchange formats specify data encoding and internal file structures that can be used to exchange data between different hardware and software applications. A useful analogy can be found in modern printers. Files sent to the printer are exported by computer programs in a format such as the Microsoft Windows Metafile (WMF) or Adobe Systems PostScript formats. WMF and PostScript are interchange formats that can be read by most printers alleviating each software developer having to write instructions for many different printers.

Every service implemented has an associated behavior model and data model. The data model is usually expressed as an interchange format. The use of interchange formats mean data producers and consumers can continue to use their internal data formats that are optimized for their business requirements. Examples might include proprietary data created by a scientific measuring device. Such data formats may be suboptimal for using as interchange formats.

A number of international efforts are under way to develop specifications for data interchange of geoscience information that are applicable to NGDS Tier 3 data types. These include (GeoSciML) (Richard and CGI Interoperability Working Group, 2007; see also http://geosciml.org), and the OGC observation and measurement model (Cox, 2010). These models are very flexible and allow representation of a wide range of content, but are thus correspondingly complex and difficult to use. Thus, in the initial phase of the project, content models have been defined using relatively simple schema in which property values are specified only by string or numeric-valued elements (no nested or complex data types). The content models are designed to be compatible with the more complex and comprehensive models mentioned above to the degree that is practical.

USGIN is currently implementing interchange formats as GML Simple Features (van den Brink et al., 2011), compatible with the service protocol in use (OGC WFS). WFS can be consumed by existing clients like ESRI ArcGIS Desktop and Quantum GIS. As clients are developed for richer-content, the NGDS can adopt more complex, information-rich interchange formats.

Versioning

Another challenge to the geothermal community is the evolution of standardized interchange formats. WFS services have been deployed using interchange formats implemented as the models evolve (an agile process), and iteration of model versions and XML schema for corresponding WFS features can easily result in discrepancies between interchange format versions.

An important part of system operation and maintenance is ongoing review of deployed services and careful validation of new services to maintain conformance with the specifications and a well thought out change mitigation plan. As part of the change management process, XML schemas are versioned and the namespace for the schema elements is unique to that schema version. Thus namespace-aware client applications can determine if an instance document is using a supported version.

METADATA

To meet its main objectives, specifically catalog-based search, discovery and retrieval of resources, the NGDS requires quality metadata describing the information resources. Metadata, in the context of NGDS, is data that describes a physical or electronic resource, provides information about the content of the resource, its origin and processing history, how the content is represented, and how the resource can be accessed. Note that the term resource is used here in a very broad sense to mean any identifiable item of interest to users of the information system.

NGDS metadata content can be generally classified into one of five categories:

Basic metadata provides information that applies to a wide spectrum of resources, and includes the title, a description of the resource, author(s) (originator), the creation (or publication) date, and specification of the natural language of resource content. Information used for metadata maintenance, such as the metadata record ID, update date, point of contact, and metadata specification name are also included in this group.

Guide metadata is used to help users find, evaluate, and access specific items. This group includes access instructions, distributor contact information, bibliographic citation, a unique identifier for the resource, links to access the resource online, keywords categorizing the resource, information about the quality of the resource content, the geographic area described by the content, and any constraints on access or usage of the resource.

Process metadata captures more specific items such as the process used to create the data, the purpose of the data and the context in which the data was created. This could include a description of machines used to measure or sample a physical entity or items like testing processes used to derive the data.

Structural metadata is used to describe the structure of data such as syntax, serialization, tables, columns and indexes. This term can also be used to describe the organization of the data and if electronic, the serialization or Multipurpose Internet Mail Extensions (MIME) type used. This may also include details of physical objects such as drill core samples. The basic NGDS metadata only includes information on the format of an online representation of the resource. The full ISO 19115 (2003, 2006) (schema includes many additional fields for describing resource structure that may be included in NGDS metadata.

Domain specific metadata is often applicable only to a particular data type, and is thus not suitable for inclusion in metadata meant to be applicable across the entire range of NGDS resources. Such information can be included in the description or lineage statement fields in the basic metadata content, if it applies to all records in a dataset. The individual content models developed for NGDS data types include fields for metadata content specific to individual data instances of that type. Metadata at the domain (feature) specific level is accessed through data services for the particular feature type, or may be understood by studying written documentation.

The NGDS adopted the minimum metadata content recommendations for geoscience resources and the metadata content and encoding profile developed by the USGIN project (USGIN, 2011a, 2011b). For NGDS catalog purposes, the USGIN recommendations have been relaxed, allowing some of the recommended mandatory fields to be nilable—i.e. the field must be populated, but a value of 'missing' or 'not applicable' is allowed to indicate that the information is not provided. In the end, the practical minimum metadata requirement is that there is an informative title, some kind of geographic location information, and sufficient information for a user to know how to get the resource. If a document or dataset is not specific to any geographic location, the location keyword 'nongeographic' is used. For geographic location, a latitude-longitude bounding box is the preferred specification, but lacking that, placename keywords are allowed. Document-based resources registered by project partners are expected to be accessible on the web, in which case the access information will be a web location (Universal Resource Locator or "URL" for short) that will get an electronic version of the document. For physical NGDS referenced artifacts, a contact point to request access to the resource should be part of the metadata.

At a more granular level, individual records (features, objects) in a dataset may include source information, documenting details of observation or measurement procedure and other information specific to a particular data type. This might include information such as location, data and time of observations, and the source of that data. These feature-level metadata are delivered with the data, and only summarized in the work-level dataset metadata that are published to the NGDS catalog. This granularity issue can be difficult because of differing perspectives on what is data or metadata, differing granularity of documentation available, and different use-case priorities.

All geo-scientific data (e.g. geology, geochemistry, geophysics, remote sensing, temperature surveys) require geographic coordinates to place the data in the proper spatial configuration for analysis. The NGDS requires data input to include surface location information, and depths where appropriate, for all data input such that users can query multiple data sets (and publications) to obtain relevant information for their analysis of either site specific or regional geothermal areas. As such, wherever possible, all data is spatially located for ease in locating and using data specific to user's needs. The data included allow work to be conducted in all phases from preliminary, geothermal exploration of regions and target/area identification to site assessments and resource development of specific areas.

Annotations

The Annotation class in Figure 2 represents tags, ratings, event log items, comments or links to other resources that are asserted by NGDS users, as opposed to the data owner, steward, or provider. The ability for users to associate annotations with NGDS resources provides a feedback mechanism resulting in an emergent knowledge base.

Tags are a kind of annotation that consist of plain language text terms assigned by users to categorize

resources according to schemes that they find useful. Because individuals think differently, it is useful to enable users to augment metadata by adding such annotation. This *bottom up*, collaborative process produces what is commonly known as a folksonomy (http://en.wikipedia.org/wiki/Folksonomy). The NGDS team believes that incorporation of such crowd-sourced tagging will help improve search effectiveness by combining this approach with the top down controlled keyword approach and using a thesaurus-like functionality.

Metadata Acquisition

The project participants have used various metadata content schemes that must be harmonized to enable an integrated catalog search. Existing metadata includes lists of files compiled in a text document or spreadsheet, various databases constructed by organizations to manage their library holdings, and formal metadata conformant in varying degrees to Federal Geothermal Data Committee (FGDC) or rarely ISO standards, constructed according to locally varying interpretation and practice. In some cases, the metadata collected is not sufficient to conform to the USGIN recommendations. Manual addition of information to complete the metadata could potentially require funding resources beyond what was budgeted.

The challenge facing the team is to minimize the manual data entry required to ensure sufficient metadata content to enable a set of use cases. Making metadata acquisition as simple as possible is a design goal of NGDS. Approaches include user-friendly forms, spreadsheet editing that is familiar to most computer users, transformation processes from existing database metadata, and automated metadata extraction. The metadata requirements were also relaxed somewhat (as noted in the metadata section) to allow 'missing' as a value for some required content.

Metadata entry workflows developed and in use for the AASG "State Contributions to NGDS" project that are contributing to NGDS include a web form interface and a spreadsheet template for compiling metadata. The form interface uses background user log-in information to auto-populate some of the metadata, as well as providing pick lists and auto-complete functions in the data entry fields. The date and timestamp of submission can be recorded, saving the data provider from having to create this data for each submission manually.

Use of the spreadsheet allows users to do extract/transform/load processing from their existing metadata table or spreadsheet using familiar cut, paste, search/replace, and fill-down operations supported by the spreadsheet software. The spreadsheet metadata compilation table columns are mapped to the USGIN ISO metadata profile, and metadata entered in each row can automatically be converted to an XML record to import into the NGDS catalog. The software that does the conversion operates on a comma-delimited text (CSV) formatted table, which can be exported from the spreadsheet software or created by a variety of other workflows.

Data providers with metadata expressed in a database schema have a variety of options for publishing the metadata to the NGDS catalog. Standardized Query Language (SQL) views that duplicate the table structure of the metadata compilation spreadsheet can be used to export CSV files that can be converted to XML. A more streamlined approach is to implement a USGIN-ISO XML export function directly against the table in the database. By saving these files in a web-accessible directory that can be harvested by the catalog, the metadata content in the database can be kept synchronized with the NGDS catalog with virtually no user intervention.

Location information

One of the major challenges for metadata acquisition is obtaining the geo-location information for the numerous resources. In order to enable the basic geographic search use cases using a map interface, each resource metadata record must have a latitude-longitude bounding box that delineates the geographic area that is the subject of the resource. The metadata creation form interface allows the user to draw a rectangular box in a map view. With care, this can produce accurate location metadata, yet this is time consuming, typically requiring 3-5 minutes per metadata record. If this is deemed too much effort, locations can be specified using place-name keywords. In some cases, if there is a good correspondence between a named location (mountain range, valley, known geothermal resource area) and the subject area for a resource, this gazetteer approach can yield good results. In many cases it may be possible to correlate the named locations with geographic bounding boxes to enable the map-based geographic search.

A large amount of geothermal data is obtained from wells that are traditionally (in the United States) located with legal descriptions based on survey bases like the Public Land Survey System (PLSS) (http://www.geocommunicator.gov/geocomm/lsis_home/home/lsis-plss-description.html). GIS datasets with the PLSS grids are available from the Bureau of Land Management for many of the western United States (http://www. geocommunicator.gov/GeoComm/lsis_ home /home/ index.htm#plss), and these enable automated mapping of consistently formatted Township-Range-Section-Quarter Section type PLSS locations to a bounding box or center point that can be used in geographic search for wells in a well header feature service.

Automatically Generated Metadata

Some metadata, such as the electronic transfer protocol used to retrieve the NGDS Data Resource (examples: FTP, HTTPS), the methods required (HTTP Post, Get) can be populated by default if the metadata is being uploaded to a repository. Structural Metadata, such as the MIME type, can be inferred during a file upload process as well.

In some cases, a file that is being registered may already contain some useful metadata. Portable Document Format (PDF) documents using version 1.5 or later include a metadata section with content defined by Extensible Metadata Platform metadata standard (XMP) (Adobe Systems, Inc., 2005). The XMP scheme extends Dublin core with a variety of properties. Recent versions of Microsoft® Office® documents also have internal metadata sections. If any of this metadata content was created with the file, a data provider may possess metadata without even realizing it. This sort of metadata be programmatically detected by the NGDS resource registration software using a software toolkit like Apache Tika (http://tika.apache.org/).

Some metadata content can be automatically generated when a resource is registered to the NGDS system. For example, an identifying Uniform Resource Identifier (URI) can be assigned automatically if none is provided, as well as the URL for accessing the resource if the file is uploaded to an NGDS node.

Manual Metadata Entry

When resources are registered in the NGDS, a metadata validation process will be run to determine that metadata requirements are met. This is necessary to ensure a minimal set of metadata to accommodate all the user interface functionality revealed by a User Centric Design (UCD) research project performed as part of the NGDS work. NGDS data resource providers will be requested to complete any missing information. In some cases, there may be several dates associated with the data that must be manually specified, such as the curating date, the creation date and in some cases references to dates or specific tests or observations. Other information about the resource might only be obtainable from the data steward and require manual entry.

For example, the NGDS metadata content recommended more entries than that required for existing metadata from the Oregon Institute of Technology Geo-Heat Center library. The information was incorporated into the compilation spreadsheet. After further refining the content, the catalog import reviewers requested some additional changes in the way Geo-Heat Center data was entered. For example, Geo-Heat Center added a column with location allowing the program to define the bounding box. Keyword entries were separated by a pipe symbol instead of semicolons.

The job of manually creating and verifying the metadata is shared among several roles as described in an NGDS Software Requirements Specification (SRS) document. This alleviates one person from an unfair burden of work and also ensures that quality checks are performed.

METADATA COMPARISON

The NGDS team has compared metadata elements from the different profiles in use to the USGIN ISO metadata profile to determine how compatible the metadata standards are with each other. The metadata models being used within NGDS included:

1. The U.S. Geoscience Information System (USGIN) Profile for ISO 19115/19139 (2003, 2006);
2. The NGDS Metadata Compilation Template v1.3.4, which is a simplified, flat-table view of the USGIN profile;
3. The National Renewable Energy Laboratory (NREL) DOE-GDR Metadata Template (Weers and Anderson, 2013);
4. Dublin Core metadata vocabulary. Dublin Core is a basic set of metadata commonly used to declare citations and associate authors and other attributes with documents. This vocabulary is used by the DOE's Office of Scientific & Technological Information (OSTI) (http://www.osti.gov/OSTI_OAIrepositorymanual.pdf) and DCAT (http://www. w3.org/ TR/vocab-dcat/);
5. Metadata terminology and taxonomies developed for geothermal data collecting and mapping by the Southern Methodist University (SMU) team as part of a heat-flow data base development project for NGDS;
6. Ordinary plain language folksonomy terminology that arose from the NGDS UCD work.

In the metadata comparison shown in Figure 4 each row has a label for the baseline metadata concept in the left column, and columns for the corresponding metadata content field labels from the schemes to be harmonized listed above. Each row represents a different metadata content element and includes the terms used for that element. The grey shaded boxes indicate places where a model includes no content item corresponding to the concept in that row. Many of the metadata content items were easily mapped to the concepts in column one.

Information in fields that do not map directly to ISO metadata elements can be included in the free text abstract field to be made available to users.

REQUIREMENTS FOR DATA ACCESS

When interviewed, users participating in the UCD study indicated that they prefer to search using a map view as the interface, and would like to know what data exists within the given boundary of a shape on the map. A typical prospector would likely start a search with a map to see the location of data acquisition sites (e.g. wells, outcrops) and access information available from those sites. The ability to filter the data based on whether or not it is within a defined map area would require that the metadata include geospatial location information. The ability to find results meeting the rest of the criteria would require accessing the actual datasets. The metadata would guide discovery and access to the appropriate datasets, which would then need to be analyzed and integrated to respond to identify the target sites.

This will facilitate the types of searches required from the UCD work. To elaborate on this, consider the following use case:

"A geothermal prospector or resource geoscientist working for a land owner or potential developer/investor has a property presented as a geothermal prospect. They need to know what geological, geophysical, land use and other datasets exist that are relevant to the decisions to be made about the viability of a geothermal project. For example, finding data that could indicate there are springs nearby, geochemical geo-thermometer data for the water, data about wells in or near the area including the depth and temperature, the nature of the heat flow gradient and the heat flow of the wells, is a benefit to geothermal prospectors. Additionally being able to locate and retrieve copies of any relevant publications (geological, geothermal, et al) that deal with the area or document it in more detail would be useful."

This requires that the metadata contains latitude and longitude coordinates in order to map them in relationship to roads and power grids. The content model for the dataset must have the temperature data and related depth information. If the data resource is a Tier 1 unstructured resource, the metadata records corresponding to that record should indicate that the data exist in a non-programmatically accessible format so that an individual will be able to review it. Additionally, the metadata for the resource must specify the web location of the NGDS Resource, the protocol used to retrieve it, the resource identifier, and a description of the resource if it is in an electronic format.

	A	D	F	H	I
1		Metadata Specifications Field List Comparison Worksheet			
2	BASELINE for comparison Technical metadata interchange content model: Metadata Compilation Template v1.3.4 (and 1.3.5) (fields reordered in groupings)	Plain English metadata capture template (derived from technical metadata interchange content model)	NREL DOE-GDR Metadata Template (Jon Weers)	DCAT Vocabulary (CKAN-supported specification)	DATA.GOV Metadata
3	Element	Field	Field	Property	Element
4	title	Data Resource Title	Data Resource Name	title	Title
5	description	Data Resource Description	Description	description	Description
6	publication_date	Publication Date	** inferred	release date ; dataset update/modification date	Date released ; Date updated
7	resource_languages	Data Resource Language	** inferred	language	
8	resource_id**	Data Resource Name**	** generated	identifier	User Generated ID
9	resource_url**	Data Resource Path**	Data Resource Filename & Path		Access point
	resource_type	Data Resource Type	** inferred	format	File format :

Figure 4. Metadata comparison spreadsheet

Without such metadata declarations, the NGDS user would not find the data

In some cases a required dataset does not come from the geothermal community. Data such as road information, power grid location and land ownership will be provided by other authorities. Using a Service Oriented Architecture (SOA) accessing map data through standard OGC web services enables mashups that include NGDS data with layers added from other sources.

UCD user-research also identified a vocabulary of keyword terms suggested by potential NGDS users to provide guidance on user-expected search methodology. The studies showed that the intended user constituency commonly searched for information using different terminology than those providing the data. In some cases, the words were simply synonyms such as "drill hole" vs. "bore hole". Mapping between the various tag and keyword vocabularies from different organizations and communities is an area of ongoing research.

Access to Large Nationwide Datasets

Facilitating detailed data discovery for highly geographically dispersed datasets is a significant user interface challenge. For example, the SMU heat flow database contains data for thousands of sites throughout the country. The catalog metadata record for this database can only indicate the kinds of information that might be available for each site, and the boundary of the region that contains all the sites. A user must interact directly with the database to learn the precise location of the sites and what information is available from a particular site. In this instance, the catalog would lead the user to a Web Feature Service (WFS), that places the data from the system into an explicitly defined structure (the Content Models referenced previously) that can then be manipulated by one or more front end applications that merge the data with other available resources, such as road information, power grid locations and water resources. The client application must enable the user to navigate from the metadata for a whole regional dataset to the detailed data from individual records in the dataset using a seamless process that requires minimal input from the user.

LESSONS LEARNED

Demonstrating the value of metadata and Tier 2 and Tier 3 data to data providers and users is an ongoing challenge. Many users might think they are referring to metadata when they are actually referring to instance data, or information about a specific field within a data record. The user interface must be carefully designed and constructed to guide users in an intuitive way. By carefully architecting the search, graphical user interfaces for the data retrieval parts of NGDS, users must be able to get the desired results without having to understand the inner working of the system. In order to facilitate this, both the data and metadata models have to be well thought out.

Good metadata is essential to the success of the system and obtaining this information must be made as simple as possible, ideally seamlessly integrated into workflow such that the user is hardly aware that they are 'creating metadata'. Automating metadata creation wherever possible is part of this philosophy. On the other hand, it is also clear that users must be involved in the process to detect errors and omissions. While some metadata can be generated and validated automatically, users should remain involved to ensure the results are both complete and accurate.

CONCLUSION

With the assistance of geothermal domain experts, metadata and Tier 3 data specifications and information exchanges are currently in production mode (see http://geothermaldata.org). Current NGDS development in progress as of the date of this paper is focused on implementation of a portal application for searching all NGDS resources, and 'Node-in-a-Box' software that will simplify deployment of new NGDS nodes and their incorporation into the system. It is a highly complex problem involving both technology and human components. NGDS teams working from both the user centric approach and the data provider side are making progress. The ultimate indicator of success will be known when the NGDS system goes live in early 2014 and real world usage patterns emerge.

A greater geothermal community of practice will emerge as data needs are addressed and the value of an interoperable network is demonstrated. Only then will geothermal community fully engage.

REFERENCES

Adobe Systems Incorporated, (2005), "XMP Specification: Adobe Systems Inc.," San Jose, CA, 112p. (Accessed at http://partners.adobe.com/public/ developer/en/ xmp/sdk/XMPspecification.pdf)

Adobe Systems Incorporated, (1999), "ISBN 0-201-37922-8," (Accessed at http://partners.adobe.com/public/developer/en/ps /PLRM.pdf)

Bloomberg, New Energy Finance, (2012-10-08), "Geothermal Market Outlook, Achilles heels: resource risk and land rights".

Clark, R., Kuhmuench, C. and Richard, S. (2013), "Developing the National Geothermal Data System; Adoption of CKAN for Domestic & International data deployment," Proceedings, Thirty-

Eighth Workshop on Geothermal Reservoir Engineering, Stanford University, Stanford, California, February 11-13, 2013, SGP-TR-198

Cox, S. editor, (2010), "Geographic Information: Observations and Measurements," OGC Abstract Specification Topic 20, v2.0.0,":Open Geospatial Consortium Inc., Document 10-004r3, same as ISO 19156. (accessed at http://portal.opengeospatial.org/files/?artifact_id =41579 2013-01-24).

De La Beaujardiere, J. editor, (2006), "OpenGIS® Web Map Server Implementation Specification," v. 1.3.0: Open Geospatial Consortium Inc., Document 06-042.

Dublin Core Initiative, (2008) "Dublin core Metadata Element Set, Version 1.1: Dublin Core Metadata Initiative," (accessed at http://dublincore.org/documents/dces/).

ISO 19115/Cor.1. (2006), "Geographic information – Metadata," Technical Corrigendum.

ISO 19115. (2003), "Geographic information – Metadata".

Kottman, C. and Reed, C., editors, (2009), "The OpenGIS® Abstract Specification Topic 5: Features, v. 5.0," Open Geospatial Consortium Inc., Document 08-126.

Portele, C. editor, (2007), "OpenGIS® Geography Markup Language (GML)," *Encoding Standard*, **3.2.1**: Open Geospatial Consortium Inc., Document 07-036.

Richard, S. M. and CGI Interoperability Working Group, (2007), "GeoSciML – A GML Application for Geoscience Information Interchange," U. S. Geological Survey Open-file Report 2007-1285, p. 47-59. (accessed at http://pubs.usgs. gov/of/2007/1285/pdf/Richard.pdf, 2013-01-24)

United States of America, Senate and House of Representatives, (2009), "American Recovery and Reinvestment Act of 2009," (accessed at http://www.gpo.gov/ fdsys/pkg/BILLS-111hr1enr/pdf/BILLS-111hr1enr.pdf).

USGIN (2011a), "Metadata Recommendations for Geoscience Resources," U.S. Geoscience Information Network, document USGIN2011-002 accessible at http://repository.usgin.org/uri_gin/usgin/dlio/335

USGIN (2011b), "Use of ISO metadata specifications to describe geoscience information resources," U.S Geoscience Information Network, document USGIN2010-09, accessible at http://repository.usgin.org/uri_gin/usgin/ dlio/337.

van den Brink, L., Portele, C. and Vretanos, P. editors, (2011), "Geography Markup Language (GML) simple features profile (with technical note)," v2.0: Open Geospatial Consortium Inc., Document 10-100r3.

Vretanos, P. A. editor, (2005), "Web Feature Service Implementation Specification v. 1.1.0," Open Geospatial Consortium Inc., Document 04-094.

Weers, J. and Anderson A. (2013), "DOE Geothermal Data Repository (GDR): Fueling Innovation and Adoption by Sharing Data on the DOE Geothermal Data Repository," Proceedings, Thirty-Eighth Workshop on Geothermal Reservoir Engineering, Stanford University, Stanford, California, February 11-13, 2013, SGP-TR-198