EERE

Guidelines for Provision and Interchange of Geothermal Data Assets

Supporting Advancement of Geothermal by Populating the DOE Geothermal Data Repository, a Node on the National Geothermal Data System

US Department of Energy Geothermal Technologies Office

January 27, 2016

Version 3.4

The National Geothermal Data System (NGDS) is a distributed information system providing access to integrated data in support of, and generated in, all phases of geothermal development. DOE's Geothermal Data Repository (GDR) is a node within the NGDS system. Data uploaded to the GDR is made available through the NGDS catalog. NGDS is built in an open paradigm (the U.S. Geoscience Information Network, USGIN, which is operated by USGIN Foundation, Inc.) and employs state-of-the-art informatics approaches and capabilities to advance the state of geothermal knowledge in the US. This document presents guidelines related to provision and interchange of data assets in the context of the National Geothermal Data System, identifies general specifications for NGDS catalog metadata and data content, and provides specific instructions to GTO-funded projects for preparation and submission of data assets to the GDR. The GDR is an open data repository that houses data from DOE funds recipients.

Contents

Introduction

The Geothermal Technologies Office (GTO) funded the development of the National Geothermal Data System (NGDS), a distributed information system providing access to integrated data in support of, and generated in, all phases of geothermal development. The NGDS increases the likelihood of success in geothermal development and reduces risk by providing a comprehensive technical and economic knowledge base. The NGDS is an environment through which all types of geothermal-relevant data and publications can be disseminated, enhancing understanding about both successes and failures in geothermal research and development and providing indicators concerning the appropriate use and quality of data. The NGDS is built using the U.S. Geoscience Information Network (USGIN, operated by USGIN Foundation, Inc.), an open paradigm supporting broad access by users of all types to data and tools relevant for their work while protecting information when required and employing state-of-the-art informatics approaches and capabilities to advance the state of geothermal knowledge in the US. GTO's Geothermal Data Repository (GDR) is DOE's node on the NGDS and is home to geothermal data from DOE funds recipients. The technical elements of this approach depend greatly upon the documentation of structured and unstructured data using standardized metadata and common data types.

This document presents guidelines related to provision and interchange of data assets in the context of the National Geothermal Data System, guidlines on submission organization and data formatting, and guidelines on submission timing, data requirements, and metadata description.

For specific guidance on uploading data to the GDR, see the help video or FAQ on the GDR (https:/gdr.openei.org), or view your project terms and conditions.

In general, NGDS metadata and submission requirements are built upon domestic and international standards[1] for information dissemination including Project Open Data (GSA), DOE's Office of Science and Technical Information (OSTI), and recommendations from the US Library of Congress's recommended formats page.

Metadata Provision and Interchange

Specification of metadata for any data asset provides the information necessary to list that asset in a searchable catalog and enable discovery through various search mechanisms. The minimum, essential information that must be provided includes a meaningful title, source attribution, information on how to get the data asset, and sufficient documentation that a future user can understand the origin and content of the data asset well enough to determine if it meets their needs and how to use it properly. Specific metadata format requirements and templates are determined by the policies of the repository to which data assets are submitted.

[1] http://dublincore.org/specifications/

General considerations when creating metadata

The metadata provided with data submissions to the NGDS allows those submissions to be added to all of the data catalogs in the NGDS. Because these submissions are propagating to multiple sites, it is important to think of submissions to the NGDS as a means of reaching a greater scientific community, as opposed to a submission to a single site. As such, all metadata fields should be completed in a manner appriate for consumption by the greater community. Do not assume domain knowledge or project familiarty when considering the intended audience of your data submission. The following field-specific guidelines are meant to assist you with creating the best metadata possible for your submission:

Title: The title should be informative and clearly indicate the nature of the resource. For instance, a title like "Pima well photos and notes" is ambiguous and leaves many questions unanswered about the named resource. Similarly, a title like “Project 42 Final Report” is meaningless to anyone not familiar with the internal project number and does not convey the nature of the information contained within the report. A title like "Photos of TKBG-3025 well site, Pima County, Arizona, and chronologic notes on drilling progress" would be much more useful.

Abstract: The abstract should summarize the content of the submission, where and how information was collected, why it was collected, and if possible provide the salient features of the data including value ranges, number of items, predominant subjects, location of the area of interest, and major geologic units described.

Source: The relevant metadata fields should be completed to include sufficient provenance information for users to understand who collected the data, how, why and when it was originally collected, and what kind of processing it has undergone.

Keywords: Sufficient keywords should be added to assist in the discoverablity of the data submitted. Keywords should reflect the who, what, where, and when of the data submitted and should be reflective of the data submitted and the abstract provided. Various DOE projects may specify particular keywords to be included in repository submissions for project products. Some project-specific provisions are included in Appendix F.

Resource-specific information: For structured data that is not in an NGDS Tier 3 model, the data dictionary for tabular data – the explanation of what the fields in the data are about, units, types, vocabularies and assumptions – should be explicit in the data. This is easily accomplished through the use of verbose column headers or supplemental tabs. If the data information is not explicitly described in the resource, then it should be included in the metadata (abstract or description) for each individual resource.

Data Provision and Interchange

For the purposes of data documentation and access in NGDS, a basic distinction is made between unstructured and structured data assets. Unstructured NGDS data assets are those data assets whose content is not easily parsed by a computer program into separate and distinct data items. These assets

are packaged in files and require a human to extract data for analysis. Common unstructured technical formats include: Adobe PDF, Microsoft Word, Microsoft PowerPoint, various image formats, various audio formats, etc. Each unstructured data asset must be registered in an NGDS catalog using sufficient metadata and must be directly accessible on the World Wide Web using, a unique HTTP URL (resource locator).

Structured NGDS data assets are those data assets whose content can be parsed into separate and distinct data items by a computer program. Common structured file formats include: XML, JSON, LAS, csv, xslx (Microsoft Excel) and MDB (Microsoft Access). Each structured data asset must be registered in an NGDS catalog with metadata that documents the structure of the data (e.g. a data dictionary), and must be either directly internet accessible (required if the asset is file-based), accessible through appropriate data access Web services, or both. Structured data sets in proprietary and/or binary data formats should only be registered if there is no other option, and their metadata must clearly indicate the specific software necessary to access the content of the files.

Tiered data delivery

The NGDS was designed using a tiered data delivery scheme to accommodate unmanaged legacy data in whatever form it is available, as well as high value data in standardized content models. The tiers are defined as follows:

- Tier 1: Unstructured Data Assets – files containing information in formats useful to human users but not designed for automated consumption.
- Tier 2: Structured data, but not standardized —data structured in formats that are not conformant with an adopted USGIN/NGDS content model. Data in this tier are amenable to machine processing, but would need to be transformed in some fashion by a data consumer to integrate with other datasets.
- Tier 3: Structured, standardized data — data published using an adopted USGIN/NGDS content model and interchange format.

For tier 3 data assets, the NGDS uses content models defined in the USGIN content model registry (http://schemas.usgin.org/models/). These models define data interchange content and encoding for commonly occurring thematic data types**,** particularly for exchanging data using Web services. Content models enable consistent exchange of data among servers and applications, thereby enhancing interoperability among computing systems. The system uses a community governance scheme to adopt new interchange formats (see Defining New Information Exchanges and Editing Existing Exchanges), and provides a repository where the specifications for each data exchange are available to all (https://github.com/usgin-models).

Tier 3 data delivery is the preferred scheme, but because of the additional effort required to edit and review datasets for Tier 3 delivery, the NGDS has developed these models for thematic data types that are common. This has been done by informally by surveying data providers to determine the types of structured data that they have available for inclusion in the NGDS, and geothermal exploration and development practitioners (mostly in the State Geological Survey community) to determine which of

these types should be prioritized. Data that is very specific to a small group of users should be delivered as Tier 2 data; some extra effort is required to document the structure of these 'custom' datasets, the meaning of table and field names, and the meaning of any specialized vocabulary used in the data.

Content models adopted by USGIN/NGDS to date include:

- Abandoned Mines
- Active Fault/Quaternary Fault
- Aqueous Chemistry
- Borehole Lithology Intercepts
- Borehole Lithology Interval Feature
- Borehole Temperature Observation
- Contour Lines
- Direct Use Feature
- Drill Stem Test Observations
- Contour (Isoline) Feature
- Fault Feature/Shear Displacement Structure
- Fluid Flux Injection and Disposal
- Geologic Contact Feature
- Geologic Reservoir
- Geologic Units
- Geothermal Area
- Geothermal Fluid Production
- Geothermal Metadata Compilation
- Geothermal Power Plant Facility
- Gravity Stations
- Heat Flow
- Heat Pump Facility
- Hydraulic Properties
- Physical Sample
- Powell and Cumming Geothermometry
- Power Plant Production
- Radiogenic Heat Production
- Rock Chemistry
- Seismic Event Hypocenter
- Thermal Conductivity Observation
- Thermal/Hot Spring Feature
- Volcanic Vents
- Well Fluid Production
- Well Header Observation
- Well Log Observation
- Well Tests

Additional content models under consideration:

- Financial
- Daily Drilling Report
- Geophysical Survey Results
- Hydraulic Property Observation
- Subsurface Alteration
- Surface Alteration
- Well Completion Information
- Well Production Hardware

Note: for any given data asset repository, if there is not an existing accepted format defined for a particular type of data, the.

All content models are available at http://schemas.usgin.org/models/. They are under change management and the content model directory will provide access to current and previous releases. Also, new content models are being added to the collection. For a current view of these and any new content models, or to suggest a new model, please see: https://github.com/usgin-models as well as documentation at https://github.com/usgin/usginspecs/wiki/. The existing documented content models may be used as templates to format data sets for which an existing model does not apply, and provide templates for defining new models.

Data Plans for DOE-Funded Projects – What to Include

For each task of the project, the following questions should be answered:

- What externally sourced data are input for this task?
  - Describe the data to be acquired (especially structure and meaning).
  - Why are they needed for this task?
  - From where are they sourced?
  - What is known about the origination of the data?
  - Is there a cost for access to the data (licensing)?
- What data will be generated as a result of the task (raw or derived)?
  - Describe the data to be produced (especially structure and meaning).
  - What is the expected value of the produced data?
    - Who is likely to use/benefit from the data?
    - How are the data likely to be used?
    - Provide an example of how the data will likely be used and by whom.
  - Is produced data raw data or derived data?
  - If derived, by what method/algorithm and what version of that method/algorithm? Proprietary algorithms should be identified but their details need not be revealed.
- What categories of geothermal-relevant data will be produced (raw or derived)?
  - See recommended data categories – Appendix A
- In what formats will datasets be produced?
  - See recommended technical formats – Appendix C
- What standards apply to production of datasets?
  - See applicable standards – Appendix D
  - See also expected metadata attributes – Appendix B
- To what NGDS node(s) will each project dataset be published?
  - For datasets from DOE-funded work the default repository is the DOE Geothermal Data Repository (DOE-GDR); alternatives must be negotiated with DOE project officers.
  - Data providers may also chose to become a node on the NGDS by downloading the free software stack “Node-in-a-Box” at http://geothermaldata.org.

Submitting Data Assets from DOE-Funded Projects

Why must data assets be submitted?

> "Our **success** should be measured not when a project is completed or an experiment concluded, but when **scientific and technical information is disseminated**. Beyond broad availability of technical reports, e-prints and multimedia, and publication in peer-reviewed journals, **open access to experimental data and analysis codes** is increasingly important in policy-relevant research areas. The Department will establish guidelines for use with both grants and contracts to **ensure appropriate access to, and retention of, scientific data and analysis methods**. In more applied areas, **knowledge of what did not work** can be of equal value with **positive results**, for that can prevent the misapplication of significant private resources. The Department will therefore encourage the documentation and archiving of negative results from all its performers **using the most advanced informatics tools**."

DOE Strategic Plan, May 2011
Assure Excellence in R&D Management, pp. 43-44

Who are the players and their roles in the data submission process?

- Submitter – preparing and submitting data assets
- DOE – Oversight, technical monitoring, funding for NGDS and DOE Geothermal Data Repository (GDR) development
- NREL – hosting and supporting the DOE GDR application and web site, assisting data providers with data curation

What data assets must be submitted?
See the recommendations in Appendices A and B for more details.

- Data, produced by a DOE-funded project, that:
    - has relevance to any aspect of geothermal development
    - could potentially be useful to others outside of the project
- Data, preferably in a "consumable" and non-proprietary format
    - Data structured to meet one of the NGDS content models are the preferred submission type
    - Internally structured (CSV, Excel, XML, mdb, dbf, LAS, LIS,) commonly used formats are preferred over unstructured (detailed below); the data dictionary should be documented in the metadata.
    - Unstructured (PDF, Word, text file, TIF…) commonly used formats are acceptable
    - Proprietary raw data formats are acceptable, only when conversion to non-proprietary formats is impossible or if such conversion incurs a loss in data fidelity. The software necessary to use the data must be clearly specified in the metadata.
    - If a particular kind of data asset is required by some government regulation, data asset formats required by regulation will be considered satisfactory for submission to DOE-GDR
- Accompanying metadata documentation for data discovery and reuse by others – tied to each data asset submitted. Metadata includes items such as:
    - Author(s) – for due credit
    - Title
    - Description
    - Publication Date
    - Geographic coordinates – for mapping purposes
    - Timeframe – begin and end date/time for activities
    - Keywords – for searching
    - Categorization – for proper filing

Where are data assets submitted?

- Data assets from DOE-funded projects are submitted to the DOE-GDR at: http://gdr.openei.org
- Data assets are preserved intact, exactly as submitted, and may be retrieved at any time by the submitter

When do data assets need to be submitted?

- Within 90 days of completion of a data set
- At termination of the project for any reason

Note: Data assets are made publicly available immediately[2] upon submission, unless the submitter negotiates a specific "moratorium" period. If such a moratorium is specified a data asset will be secured in a protected data store until the moratorium expires, at which point the asset will become publicly available.

How do data assets get submitted?

- See the instructions at: http://energy.gov/eere/geothermal/data-provision-instructions-all-doe-geothermal-technologies-office-funds-recipients

"Nodes" in the National Geothermal Data System (NGDS)

The NGDS is a network of nodes, each of which is a web-accessible server that hosts at least one of the functional capabilities enumerated below to play a role in making geothermal-relevant data assets accessible. To be considered a node in the National Geothermal Data System, the following criteria must be met:

1) The capabilities offered MUST play a role in making geothermal-relevant data assets accessible.
2) The capabilities offered MUST be publicly Web-accessible.
3) A node MUST offer at least one of the following capabilities:
   a) Host a Web-accessible repository of geothermal-relevant data assets with metadata conforming to the USGIN ISO profile (USGIN Standards and Protocols Drafting Team, 2010-11) published through an NGDS catalog node.
   b) Host a web-accessible folder that 1) contains NGDS-conformant metadata files and 2) is registered for harvesting by an NGDS catalog node.
   c) Host NGDS-conformant web services (WMS, WFS, etc.) that are registered in an NGDS catalog.
   d) Host an NGDS catalog node, which is a server operating a CSW 2.0.2 service that offers metadata conforming to the USGIN ISO profile.

A node is a server that provides NGDS resources, not a client that uses the offered resources. A server that hosts applications utilizing NGDS services is a vital part of the system, but is not considered a node. Such applications would be expected to be registered in the system by submitting metadata to an NGDS catalog.

Each node is under the stewardship of a data provider. Individual organizations may host more than one node. For example different servers may host OGC web services, a CSW service, and a repository containing document resources, all under the stewardship of a single data provider. These would be considered three distinct nodes. A node may host data from more than one data contributor.

[2] a submitter may request – without negotiation– that public availability be postponed for a year

If a server is hosting resources that are described by metadata in an NGDS catalog, and those resources meet the 'geothermal-relevant' criteria, then the operator of that node MAY declare it to be an NGDS node.

Functional NGDS node categories

Both data resources and metadata can be made accessible either packaged in files or through services that offer capabilities extending HTTP. Thus NGDS nodes can be categorized according to whether they present information in files or through web services. An individual node my offer both kinds of information access. NGDS nodes can be categorized functionally based on the access methods and the kinds of information a offered by the node.

Files-only node

A publish-only node hosts web accessible files that are either information resources registered in an NGDS catalog, or NGDS-conformant metadata records in a web-accessible folder registered for harvest to an NGDS catalog.

Service node

A service node offers at least one NGDS service that is registered in an NGDS catalog, either through the metadata for datasets exposed by the service, or as a catalog service. These are functional Web-based programming interfaces that enhance interoperability among the nodes that comprise NGDS. Instead of directly accessing data repository content, end user applications access NGDS content through these data access Web services.

Catalog Node

A service node that offers an Open Geospatial Consortium Catalog Service (CSW) to search a metadata collection hosted by that node. Metadata from catalog nodes will be harvested to the NGDS aggregator at http://search.geothermaldata.org. Like a library catalog an NGDS catalog lists the data assets available in repositories covered by the catalog, captures metadata describing those data assets, provides methods for discovering relevant data assets, and indicates the means for acquiring a data asset once discovered.

Data node

A service node that offers at least one registered data service (e.g. WFS, WMS, WCS, OpenDAP).

Publisher

Any node that makes resources (data or metadata) available that originate from the node provider.

Aggregator

A node that collects (harvests) metadata or data from other nodes for indexing and to provide value-added services, such as a catalog service, visualization, data analysis.

Data Server

A node that provides access to resources, either as files or via web services.

Appendix A. Project specific recommendations

A.1. FORGE

GDR submissions related to FORGE projects should include the keywords "EGS" and "FORGE" in their metadata. Other keywords that should be included are:

- Site Name (e.g. Fallon, Newberry, Milford)
- Project number
- Data type category. Use terms from Attachment 3 in *FORGE FOA-0000890_amd 000004*
    - Thermal Characteristics
    - Rervoir and subsurface characterization
    - Surface Data
    - Geologic
    - Existing wells
    - *More specific terms under these headings in Attachment 3 may be used in addition.*

A.2. Play Fairway Analysis (PFA)

GDR submissions related to PFA projects should include the keywords "PFA" and "Play Fairway" in their metadata. Other keywords that should be included are:

- Site Name (e.g. Fallon, Newberry, Milford)
- Project number

A.3. Strategic Materials / Rare Earth Elements

GDR submissions related to strategic materials projects should include the keywords "strategic materials" and "rare earth elements" in their metadata.

Appendix B. Recommendations for specific data types

B.1. Gridded Data

Many datasets representing continuously varying state variables, such as gravity or magnetic field variations, are encoded as grid datasets, also known as coverages or raster datasets. In this approach, an array of cells is defined that partitions the domain of interest (e.g. a geographic area or 3-D volume within the Earth, an interval of time, the trace of a borehole or ship track), and values for the variable are recorded for each cell. In most cases, a systematic scheme is defined to define intervals on each dimension of the domain space that bound each cell that has a recorded value. For example a dataset representing wind velocity as a function of space and time may be defined with a UTM grid dividing geographic space into 100 m intervals on the northing and easting axis, and time into 1 hour intervals. The origin and number of intervals on each of these dimensions must be specified as well to fully characterize the grid of cells. A pattern for matching a linear array of wind velocity values (e.g. northing,

then easting, then time) then can be define allowing a linear array of wind velocity vector values to be mapped to the space-time domain of grid cells.

Commonly used formats for this kind of data include the open-source NetCDF and HDF data formats. ESRI software has a 'grid' format that is widely used in the GIS software community, and various text based approaches are also used. Metadata for gridded data should describe the dimensions (axes) of the grid—what they represent and how the bounds of grid cells on that dimension are defined, and the kind of variable that is assigned a value at each grid cell. The variable may be single valued (e.g. temperature or magnetic field intensity), or a set of values (e.g. a multispectral image with values as several wavelengths for each cell in a raster image grid).

File formats for grid/raster/coverage data:

- ESRI ArcInfo ASCII Grid (http://www.digitalpreservation.gov/formats/fdd/fdd000421.shtml); this is an ASCII text format for a simple 2-D grid with a single value associated with each grid cell.
- GeoTiff (http://www.digitalpreservation.gov/formats/fdd/fdd000279.shtml). Format for georeferenced and geocoding information in a TIFF 6.0 compliant raster file by tying a raster image to a known model space or map projection. A GeoTIFF file is a TIFF 6.0 [TIFF_6] file, and inherits the file structure as described in the corresponding portion of the TIFF spec.
- NetCDF 4.0 (http://www.digitalpreservation.gov/formats/fdd/fdd000332.shtml). A very general format for encoding complex, multidimensional, multi-variable gridded data. A subset of HDF5. Widely supported for scientific data.

If necessary other formats may be used. Formats that are not supported by GDAL, a widely used, open-source software library for handling raster-based data in GIS systems, is strongly discouraged. A list of GDAL-supported formats is available at http://www.gdal.org/formats_list.html.

B.2. Workspace packages, data collections

Decision-making and analytical processes are commonly supported by a map-based workspace that includes various datasets organized for visualization in a particular portrayal scheme. If there is a requirement to archive data precisely as it was used in some analytical or decision making process, the recommendation is to create a repository item that records the structure, organization, and portrayal of workspace content in a single bundle. GIS software such as Quantum GIS, ESRI ArcMap, and MapInfo all define application-specific project file types that record this information, and similar workspaces are defined by other software systems for non-geospatial data.

In standard geographic information system (GIS) usage, such a workspace would consist of an ordered stack of layers in which each layer displays data from a single data source using a particular portrayal scheme. The same data may be displayed in more than one layer, using filters to display only a subset of the data in that layer, or applying different portrayal schemes to visualize different aspects of the data (e.g. heat flow points may be colored according to the heat flow magnitude, or according to the heat flow determination procedure used). In addition, the package may include a set of graphical annotation objects for the map collar (e.g. a legend, north arrow, title, explanatory text), or for labeling features in

the map visualization. The package might include complex rules for the display of labels or data depending on the scale of the current view frame. In some cases scripts or other software might be included that executes analytical processes.

Several approaches are available to record the details of such a visualization for geospatial data. The Open Geospatial Consortium has recently adopted a specification for a 'Web services context document' but this approach has yet to be widely implemented in off the shelf software; an older specification for a 'Web map context' is more widely used; ESRI provides capability to export a workspace resource from an ArcMap project that includes the ESRI map package (.mpk) file and a zip archive in which the individual data files are packed. An example tutorial on how to do this is here. Whatever approach is used, a portable workspace suitable for archiving should contain data that is in a vendor neutral format (e.g. ESRI shape file, GeoTiff…).

Careful consideration is required to determine which datasets that are used in such a workspace should be deposited in the repository, and which may simply be linked to the workspace. If any updating or modification of content has occurred, then the data should be considered a new, derivative resource that should also have a repository entry. Such updates might include, for example, clipping of polygons or lines to a study area extent, deletion of records that are considered 'bad data' for some reason, updates to correct values in fields within a data record, or merging of data from different sources. Simply selecting a subset of a larger data set is not considered to update that data, as long as the query used to select the records that are used is documented and included in the metadata for the package, and the precise version of the source dataset that was subset is accessible in a long term archive. For example a point data set of well headers might be subset to include only wells within a bounding box around a particular site, with no further update to the selected record set.

An archival repository item for a map layout package would consist of one or more 'map context container' objects. Each of these would contain a collection of datasets, as well as a 'map context document' (e.g. ESRI .mxd file) that specifies the stacking of layers, and portrayal rules used for each layer. Portrayal information might also be includes in separate files (e.g. ESRI .lyr, or OGC Styled Layer Descriptor files). The map package repository items should include browse graphics of the complete map visualization, ideally on low-resolution thumbnail, and a high resolution version in a portable format (Adobe Acrobat, .png, .jpg…).

The repository item for the package should also document each of the datasets included in the package. If the datasets are simple extracts (no updates) of existing data, then this metadata should consist of a link to the source dataset and their full metadata, and a description of the filter/query used to extract the subset used in the map package. If the datasets are updates of existing data, or consist of data assembled from various sources, these should be archived as separate repository items that contain only that dataset, and the map package metadata would include a pointer to each repository item for the exact dataset used in the map package. This procedure will result in some duplication of data, but allows for newer versions of the component data to be added to a repository item (if that is within the policy of the repository), as well as provide the necessary documentation for each component dataset; see the discussion associated with the datasets, below.

Component datasets in a map layout package:

For each component dataset included in the map package, a determination must be made as to whether the data are a new resource or a simple subset or imported resource. If the data included are accessible online and can be duplicated based on a simple select or filter type query they are considered a simple subset, and the package should just include a pointer to the source data and complete specification of the queries used to extract the subset. If an existing dataset is used in whole, with no update, then it is an imported resource, and the package metadata should include a pointer to the imported data source and to the metadata for the imported data. If metadata do not exist for the imported data, then a new metadata record must be created either in the GDR or on another NGDS node, and this will be the target of the metadata pointer.

Metadata for each resource (repository item or included subresource) should include the information outlined in Simple Metadata Recommendations for Geoscience Resources. The metadata content required by GDR is only slightly structured, so most of this content should be included in the description narrative, or as a separate Repository Item resource file containing the metadata encoded using one of the standard metadata schemes (ISO 19115/19139, CSDGM, EML).

Structured data should include documentation of the tables and fields included in the dataset, with an explanation of the meaning and data types expected for each field. Data that are loaded into table using NGDS Tier 3 content models need only state the content model and version used by the dataset to provide this documentation, and the file format used (csv, Excel, XML). If the data use some other existing interchange format (encoding and information model), pointers to documentation for that format can be provided, otherwise a data dictionary should be provided with the repository item.

B.3. 3-D models

Introduction of 3-D models into NGDS is a new development, and more information and experience will be necessary to determine best practices. Various software packages are in use to generate 3-D geologic models, but standard interchange formats for models have not been adopted. The National Archives indicates several 'preferred' or 'acceptable' formats for Computer Aided Design (CAD) vector graphics that are likely to be suitable. The degree of adoption and support for these formats by geologic modelling software vendors is not known.

In the near term, the recommended course of action is to:

1. Upload models in the file format native to the software in use,
2. Upload a copy of the model in an export/interchange format if any such format is available in the software package
3. A viewable and manipulable version of the model should be uploaded if there is an easily accessible, free software package (with a link or description of how to access it) that can be used to view the model. This may utilize the format from items 1 or 2 above, or be another format.
4. Metadata for model should contain references/links to each dataset included in the model along with description of how the data subset was selected (if applicable). The details of the file format versions and software necessary to use the files should be clearly documented.

5. High resolution (300 dpi) images should be included as well, providing visualization of the model from useful perspectives.

Appendix C. Metadata Categories and Attributes[3]

- Source
  - Author
  - Affiliation
- Quality
  - Subjective score
  - Number of flagged values
  - Percentage of missing values
- Time/Date
  - Time of measurement
  - Time of last change
  - Time of submission
- Geographic/Location Data
  - Datum (SRS)
  - Lat/Long (potentially a range of coordinates to express an area vs. a specific point)
    - Surface
      - Lat/Long Coordinates at Surface
      - Elevation of Surface Coordinates
    - At Depth
      - Lat/Long Coordinates at Depth
      - Elevation of Coordinates at Depth (Negative #)
      - Elevation of Seafloor at Depth (Negative #)
  - Legal or Map Descriptors
    - District
    - County
    - State
    - Section
    - Survey
    - Block
    - Abstract
    - Township/Range
    - Plot
    - Tract
    - Name
    - Current
    - Prior
    - Region/Area
    - UTM, etc.

[3] Siemens Corporate Research, "GTDA DATA REQUIREMENTS SPECIFICATION, Version 2.1," 07/8/2010, Appendix C, pp. 31-32.

Appendix D. Technical Format Types

- Web Sites (.html, .aspx, …)
- Files
  - Unstructured data files
    - Documents (.doc, .pdf, .ppt, …)
    - Images (.tif, .jpg, .png, …)
    - Audio (.mp3, .wav, .wma, …)
    - Video (.mp4, .mov, .wmv …)
  - Structured data set files
    - Generic structure data sets (.xls, .xml, .json, …)
    - GIS (maps, shape files, features, …)
    - Non-GIS specialized data sets (.las, lis, …)
- Databases
  - Traditional Relational (Access, SQL, Oracle, …)
  - non-Traditional (HBase, Virtuoso, MongoDB, ...)
- Software Applications (Data Access, Analysis, Modeling, Data Submittal, etc.)

Appendix E. Standards for Metadata and Data Representation

- ISO
  - ISO 19115/19119/19139 (metadata)
- Dublin Core
  - Metadata Specification
  - DC-RDF
- Federal Geographic Data Committee (FGDC)
- Open Geospatial Consortium (OGC) – Web API Protocols
  - Catalog Service for the Web (CSW)
  - Web Map Service (WMS)
  - Web Feature Service (WFS)
  - Web Coverage Service (WCS)
- HDF Group
  - HDF5
- NetCDF

Appendix F. Topical Data Categories and Attributes[4]

The following data types are common data types that may be created during geothermal exploration or surface/subsurface exploration and research. Data collected during the following research activities are suggested for submission to the NGDS. To aide in data submission, highlighted data types can be submitted in NGDS content models.

F.1. Engineering Data

- Engineering Information
- Technology
- Engineering Modeling
  - Reservoir Modeling
    - Heat Loss
    - Tracers
- Engineering Designs

F.2. Well Data

- Well Type (s) – Could be more than one, potentially
  - Water
  - Oil
  - Disposal
  - Gas
  - Injection
  - Geothermal
  - Slim Hole
  - Monitoring
  - Test
- Well Status

[4] Siemens Corporate Research, "GTDA DATA REQUIREMENTS SPECIFICATION, Version 2.1," 07/8/2010, Appendix B, pp. 25-31. (includes minor terminology corrections)

    - Producing
    - Plugged/Shut In
        - Plug Date
        - Plug Depth
        - Casing Left
        - Nearest Water Zone
        - Mud Filled
        - Number of Plugs
        - Size of Plugs
    - Temporarily Abandoned
- Depth
    - Production
        - Interval 1
        - Interval 2
        - Interval 3
    - Injection/Disposal
    - Elevation
        - Sea Level (±)
        - DF (Drill Floor)
        - KB (Kelly Bushing)
    - Vertical Depth
    - Well Bore Length
    - Bottom Hole Location (lat/long - for deviated boreholes)
- Casing
    - Size
    - Weight
    - Depth
    - Amount of Cement
    - Type
    - Date
- Tubing
    - Size
    - Depth
    - Date Installed
- Liners
    - Size
    - Depth
    - Type
    - Date Installed
- Drilling Segments
    - Size
    - Bit Size
    - Direction(s) (could be multiple directions at different depths)
    - Mud Weight
        - Density
        - Viscosity
    - Date Commenced
- Identifiers
    - API (or other identifier if API does not apply)
    - Sidetrack Identification
    - Well Number
    - Name
    - Field

- Drilling Contractor
- Logs Run (there are many possible well log types; the following are examples)
  - Nuclear Magnetic
  - Temperature Log
  - Resistivity
    - Compensated Dual Resistivity
    - Micro Resistivity
  - Induction
  - Gamma Ray
  - Neutron
  - Electromagnetic
  - Dual Spacing Neutron
  - Porosity Log
  - Spontaneous Potential

F.3. Business Data

- Financial Information (content model in development)
  - Construction costs
    - Pre-Feasibility
    - Feasibility Studies and/or Pilots
    - Land Costs
      - Drilling Site
      - Power Plant Site
    - Drilling Costs
      - Exploratory Drilling
      - Production Drilling
    - Equipment Costs (Capital)
      - Drilling Site
      - Power Plant Site
    - Electric Transmission Costs
    - Employment – Construction Phase
    - Legal and/or Regulatory Construction Costs
  - Ongoing Operational Costs
    - Equipment Costs (Maintenance, Repair, etc.)
      - Drilling Site
      - Power Plant Site
    - Employment – Operational Phase
    - Legal and/or Regulatory Ongoing Costs
  - Consumer Cost / Pricing Data
  - Insurance Costs
    - Liability
    - Financial/Loan insurance
  - Financing Arrangements
    - Government Cost Sharing (Non Tax)
    - Built to Own, Operate, Transfer, etc.
    - Loans
    - Bonds
    - Joint Ventures
- Regulatory
  - Environmental
    - Geological (mudslides, earthquakes, etc.)
    - Emissions and/or Pollution (Air and/or Water)

- Impact Analysis
- Mitigation
- Compliance

- Protected Species or Areas (Plant, animal, marine, wetlands, population centers, etc.)
  - Impact Analysis
  - Mitigation
  - Compliance

- Permits
  - Drilling
  - Transmission Facilities
  - Power Plant

- Well Owner Information
  - Rights
    - Surface
    - Mineral
    - Water
    - Geothermal
  - Oil/Gas Field
    - Name
    - Operator
  - Lease
    - Owner
    - Terms
- Taxes
  - Energy Taxes
  - Property Taxes
  - Tax Incentives
    - State
    - Federal
    - Local
- RFP/PPA/Demand

F.4. Geological Data

- Stratigraphy
  - Depth
  - Lithology
    - Induration
    - Mineralogy
      - Framework Mineralogy
      - Cement Mineralogy
    - Grain Size
  - Formation Name
- Hydrology
  - Saturation
  - Depth to Water Table
- Well Cores and/or Cuttings
  - Where stored
  - Images
  - Depth Ranges
  - Diameter
    - Full Diameter

        - Side wall
- Pressure
    - Bottom Hole
    - Wellhead
    - Shut In
    - Injection/Disposal
- Porosity
    - Depth
    - Source
    - Type
- Permeability
    - Depth
    - Method Used
    - Direction
        - Horizontal
        - Vertical
        - Unspecified
- Density
- Structure
    - Reservoir
        - Enthalpy
        - Pore Volume
        - Size
        - Thickness
    - Faults

F.5. Geochemical

- Chemistry
    - Na
    - K
    - PH Level
    - Hardness
    - Alkalinity
    - Ca
    - Mg
    - Other Elements
- Gas Saturation
- Total Dissolved Solids
    - TDS Measured
    - TDS Calculated
- Medium Sampled
- Collection Method
- Type of material
    - rock
    - fluid
- Conductivity
- Turbidity
- Specific Gravity
- Isotope

F.6. Geothermal Data

- Heat Flow

- Corrections Applied
  - Climate
  - Terrain
  - None (Uncorrected)
- Calculation Type
  - Averaged
  - Interval
- Quality Rating

- Heat Generation/Production
  - Element Content
    - Thorium (Th)
    - Uranium (235U, 238U)
    - Potassium (40K)
  - Calculated Value
- Thermal Conductivity
  - Measurement Type
    - QTM
    - Needle Probe
    - Divided Bar
    - Estimated
  - Sample Depth Interval
  - Number of Samples used to Calculate
- Temperature at Depth
  - Measurement Type
    - Continuous
      - Equilibrium
      - Non-Equilibrium
    - Discrete
      - BHT
  - Corrected/Uncorrected
  - Date/Time

F.7. Infrastructure

- Transmission
  - Type
  - Capacity
  - Owner(s) and/or Operator(s)
- Land Classification
  - Private
  - Federal
  - State
  - Adjacent
- Roads
- Power Plants – get more specific to geothermal – see examples below
  - Types
    - Nuclear
    - Geothermal
      - flash steam
      - binary
      - hybrid
      - other
    - Wind

        - Coal
        - Natural Gas
        - Biomass
        - Hydro
        - Others
    - Capacity
    - Owner(s) and/or Operator(s)
- Pipelines
    - Type
    - Distance

F.8. Production/Disposal/Injection Data

- Oil
- Gas
- Water/Brine
- Radioactive Isotopes
- Condensate
- Other
- CO2
- Geothermal
- Flow Rate

F.9. Geophysical & Synthetic Data

- Seismic Reflection Survey
    - 2D vs. 3D
- Remote Sensing
- Gravity
- Magnetics
- Geodetic
- Stress
    - Method Used
    - Magnitudes
    - Orientation
        - Frequency
        - Spacing
    - Stress Axis
- Seismicity
    - Source
        - Natural
        - Induced
            - 'Propping'
            - Production
            - Injection
    - Magnitude

F.10. Direct Use

- Applications
- Equipment/Materials

Appendix G. GDR Data Submission Guidance

Regarding the provision of data to the GDR, funds recipients are to comply with the Federal Assistance Reporting Checklist (FARC) Attachment 2, IV Other Reporting "I" Special Instructions. An excerpt from the FARC for an actual award follows:

> *DOE requires that the Recipient maintain an updated Data Management Plan containing the suggested elements for a Data Management Plan described in Appendix E of the FOA as well as information about their project data types and sources, content and format, and sharing and preservation plans by project task. As required by the Funding Opportunity Announcement for this Financial Assistance award, the Recipient must provide data to the DOE Geothermal Data Repository (DOE-GDR). The Recipient must provide data to the DOE-GDR no later than the end of the quarter in which the data are generated. The data must be sufficiently complete, in a format acceptable to DOE, and include all files required for an independent analyst to reproduce and verify the work. The data will be submitted to DOE-GDR at https://gdr.openei.org. While most data formats may be uploaded to the DOE-GDR, DOE prefers reusable, structured data that supports conclusions communicated in project quarterly and other reports. Use of National Geothermal Data System (NGDS) data exchange models is highly encouraged. For example, data generated from rare earth element analysis may be entered into a NGDS aqueous chemistry structured format available at http://schemas.usgin.org/models. The data will be made publicly available via the NGDS once they have been submitted and accepted into the DOE-GDR system. If the data are protected or subject to a moratorium, they will not be made publicly available until the moratorium has expired, and they will be held in a secure section of the DOE-GDR. Protected Data will be treated according to the Intellectual Property Provisions. Please refer to the Provision entitled "DOE GEOTHERMAL DATA REPOSITORY (DOE-GDR)" in the award Special Terms and Conditions for specific data submission instructions.*

An excerpt from an award Special Terms and Conditions follow:

> *DOE GEOTHERMAL DATA REPOSITORY (DOE-GDR)*
>
> *As required by the Funding Opportunity Announcement for this Financial Assistance award, you, the Recipient, must submit data that is generated by work on this Financial Assistance award to the DOE Geothermal Data Repository (DOE-GDR). The data will be made publicly available via the National Geothermal Data System (NGDS) once it has been submitted and accepted into the DOE-GDR, only if specified at the time of submission. If the data is protected or subject to a moratorium, it will not be made publicly available until the moratorium has expired, and it will be held in a secure section of the DOE-GDR. The Recipient shall follow the steps listed below for data submission.*

G.1. Registration

a) You must register for an account at the data submission site prior to submitting any data. Registration establishes a user account with ID and password, and authorization to submit data to the DOE-GDR. Visit the DOE-GDR website at

https://gdr.openei.org to initiate your account registration and perform all actions associated with data submission. For technical assistance with the registration or data submission interface, contact GDR Help at GDRHelp@ee.doe.gov.

G.2. Data Submission

a) Once registered, you must log into the data submission site at https://gdr.openei.org to submit data. You must complete the data submission form for each data resource (e.g. excel file, word document, pdf, or data containment software), including:
   - Provide appropriate metadata and contact information
   - Agree to the data handling terms of the DOE-GDR
   - Specify the release date for any Protected Data (if applicable), consistent with your Intellectual Property Provisions
   - Attach the data

b) After your data has been submitted, you will not be able to edit it for the duration of the review and curation process. It is recommended that you retain a copy of the submitted data.

c) For information and assistance concerning preparation of data files, metadata, unique data requirements, and the data curation process, contact GDR Help at GDRHelp@ee.doe.gov or visit https://gdr.openei.org.

G.3. Protected Data

a) Data submitted to the DOE-GDR and identified as "Protected Data" are subject to the terms and conditions set forth in your Intellectual Property Provisions incorporated into your Award. During the period prior to the public release date, Protected Data are held in a secure data store with restricted access pursuant to the Intellectual Property Provisions. All other submitted data will be made publically available once accepted into the DOE-GDR system.

G.4. Cancellation or Resubmission

a) You may cancel a submission at any time prior to public release. Cancellation will terminate the curation process and remove any copies of the originally submitted data from the system. If you wish to edit data or metadata after submission, you will need to cancel and resubmit.

b) Please refer to the following website for additional information regarding this process: http://www1.eere.energy.gov/geothermal/gdr_instructions.html