NGDS Content Models

A Tutorial

4/24/2013

Table of Contents

Section 1: Introduction

If you are a subrecipient under the Arizona Geological Survey (AZGS) for the Department of Energy (DOE) National Geothermal Data System (a subset of which is the AASG Geothermal Data project), and if you have data to submit, then you will be using the schemas provided by NGDS content models to structure your data for delivery. This tutorial will describe the usage of NGDS content models.

Content models represent one method by which the National Geothermal Data System facilitates interoperability. Content models provide a schema, or structure, for submitted data. Schemas dictate where and how data should be entered. Content models use *templates* to make it easy for data providers to submit their data in accordance with schemas.

Note: Though the terms *content model* and *template* are often used interchangeably in NGDS and USGIN documentation, a subtle distinction exists. NGDS content models are described in Excel workbooks containing multiple spreadsheets, one of which is the template spreadsheet that facilitates schema mapping. In some cases these Excel workbooks (the content models) are referred to as 'template files'; bear in mind that the template is only one spreadsheet in the workbook.

A YouTube video containing an abbreviated version of this tutorial is available here, at the NGDS YouTube channel.

Section 2: Getting Started With Content Models

Content models are used to facilitate interoperability by making it easy for computers to analyze data from multiple sources and multiple people without extensive interpretation. Consequently, data submitted for the National Geothermal Data System (NGDS) should conform to NGDS content models.

In order for subrecipients to submit data for NGDS via content models, subrecipients must first have data they wish to submit. In addition, this data must be accommodated by the content models. For example, there are no NGDS content models capable of supporting meteorological data. A list of content models can be found on the USGIN Schemas website.

If you are a subrecipient under the Arizona Geological Survey for the AASG Geothermal Data project, and if you have data to submit, the content model workflow is as follows:

1. If you do not have a software application that is compatible with Microsoft Excel spreadsheets, you must acquire one (see Table 1)
2. Go to the USGIN Schemas website and examine the content models available there
3. Download a content model that accommodates your data; if none of the content models accommodate your data
4. Assess the manner in which your data will be mapped to accommodate the content model in accordance with **Step 2** of the USGIN Data Provider Workflow.
5. Map your data to the content model by populating the **Template** spreadsheet of the content model with data, mapping your data to the appropriate field

6. Submit your data for approval by the Arizona Geological Survey; if any changes are necessary, your dataset will be returned to you; implement the prescribed changes and resubmit your data
7. Host your data using either the appropriate AASG Geothermal Data hub or via contract-designated computing resources

AASG Geothermal Data content models are stored and submitted as Excel spreadsheets with the .xls file extension. Consequently, data providers must use a program capable of reading and writing Excel spreadsheets (Table 1) to map their data to the desired content model.

Application	Developer	Compatible Operating Systems	Distribution
Excel	Microsoft	Windows, Mac OSX	Proprietary
Gnumeric	GNOME Office Team	Windows, Mac OSX, Linux, BSD, Unix	Free and open-source (GNU GPL)
Numbers	Apple, Inc	Mac OSX	Proprietary
OpenOffice Calc	Oracle Corporation	Windows, Mac OSX, Linux, BSD, Unix	Free and open-source (GNU LGPL)

Table 1: Popular, actively maintained spreadsheet applications capable of reading and writing Excel spreadsheets

Assuming you are a subrecipient with data to submit; assuming your data can be accommodated by one of the content models listed at the bottom of the page; and assuming that you have an application capable of interacting with Excel spreadsheets, the next step is to locate and download an appropriate content model.

Section 3: Finding Content Models

NGDS content models can be located on the USGIN Schemas website.

1. Using your preferred web browser, go to the USGIN Schemas website (http://schemas.usgin.org/home/)
2. Click the **View Model Details** button ()
3. Scroll to the desired content model or click the corresponding entry on the list on the left side of the web page
    a. To see a list of the fields in a given content model, click **> Field List** at the bottom of each content model description
    b. The current version of a given content model is visible on a colored tab on the right side of the screen

4. Click **Excel Template (.xls)** () to download the desired Excel workbook

Section 4: Working in your Content Model

Having found and downloaded a content model that meets your needs, find the content model on your computer and open it. The content model may be compressed as a ZIP file; if so, extract the content model from the ZIP file before opening it.

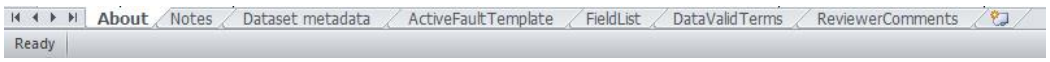

Figure 1: Tabs in Excel; each tab indicates an individual spreadsheet

Each content model is an Excel workbook containing several spreadsheets; each spreadsheet is associated with its own tab (Figure 1); each tab is described in the following list:

- **About**: Describes the content model, its authors, version information and data history
- **Notes**: Information specific to a particular content model, including conventions, color codes, and specific terminology. Notes can also be used by subrecipients to add information that pertains to the submitted dataset
- **Dataset Metadata**: Information used to populate a metadata record describing the completed dataset. Once the content model is populated with actual data, the metadata describing that data will be entered into the Dataset Metadata spreadsheet, from which an independent metadata record will eventually be generated. This metadata record will be entered into the NGDS Catalog and used to locate and access your dataset.
- **Template**: The spreadsheet in which actual data will be entered (Figure 2) according to the practices defined in the **Field List** and **Data Valid Terms** spreadsheets; for more information, see Section 5 of this tutorial

	A	B	C	D	E	F	G	H	I	J	K	L
1	HeaderURI	WellName	APINo	OtherID	OtherName	BoreholeName	ParentWellURI	Operator	MineralOwner	Driller	LeaseName	SpudDate
2		Bozo 1 OilCompany. *This row contains example data for reference only.*	42-501-20130-03-00	AzgsPer:04-004 | USGSMRD:26498-5673				OilCompany				1980-01-01T00:00
3												
4												
5												
6												
7												

*Figure 2: The **Template** spreadsheet. The cells are the top of each column contain information used to generate database fields*

- The first row of information in the **Template** spreadsheet is used to generate fields in a database. Consequently, each column represents a database field, and columns in the **Template** spreadsheet are often referred to as fields
- Each column is intended to record only the data indicated in the cell at the top of the column. For example, if the cell at the top of column D is labeled APINo, then the column should contain only API numbers for each database record. For specific instructions regarding each field, see the **Field List** spreadsheet

- It should be noted that the content models often refer to columns or fields as elements or *exchange elements*. Don't be alarmed! The terms *field* and *element* are often used interchangeably because they refer to different implementations of the same concept. *An element can be considered the XML equivalent of a database field or spreadsheet column*

- **Field List**: The **Field List** tab describes each field of the **Template** spreadsheet in detail (Figure 3), indicating the kind of data each field should contain and the manner in which that information should be entered; this description includes whether or not a subrecipient is required to populate a given field. The **Field List** will also indicate whether or not specific vocabulary from the **Data Valid Terms** tab should be used in a given field. It should be noted that not all **Field List** spreadsheets are at the same level of development – some will provide more information than others. Optimally, the **Field List** spreadsheet will provide a description for each field, as well as data entry instructions, notes explaining the peculiarities of a given field, and examples of data entered into the field.

Interchange content element	DataTypeName	Cardinality	Element Description	Element Instructions	Element Notes
HeaderURI	URI	1	Unique identifier for the borehole described by the elements in this table.	Each borehole should have a unique HeaderURI. If an APINo is present, it should be incorporated into the URI; otherwise, this URI should incorporate one of the OtherID values. Best practice is to define an http URI that will dereference to a normative description of the borehole. If possible, the terminal element of the URI should include the APINo. If the API number indicates multiple instances of reentry or recompletion, each instance is to have its own HeaderURI. For example, the API number 42-501-20130-0300 indicates a well with three sidetracks, for a total of four total boreholes, each of which should have its own database record and HeaderURI.	The HeaderURI identifies a *borehole;* wells may be constituted by multiple boreholes. This is reflected in the ParentWellURI, a URI shared by boreholes that are part of the same well. Etymological Note: the terms *borehole* and *wellbore* are often used interchangeably in USGIN - AASG geothermal data documentation. Technical Note: The term 'header record' can be used to describe a physical document produced by drilling agencies; this document logs specific instances of drilling.
WellName	alphanumeric	1	The human-intelligible name for the well containing the borehole identified by the HeaderURI.	Convention for wellName is: Name = Operator + Well No. + Lease (Mineral) Owner.	Wells may be constituted by multiple boreholes. Consequently, it is possible for multiple boreholes to be associated with the same WellName.
APINo	alphanumeric	1..n	API number or URI for the well containing the borehole identified by the HeaderURI.	Enter API number here. If an API number is unavailable, enter one or more URI(s) here. Delimit multiple URI(s) with the pipe character '|'	API number refers to the American Petroleum Institute number. Each oil, gas, or geothermal well drilled in the United States has a unique API number; this number identifies all boreholes or recompletions associated with the original wellbore. See the Data Valid Terms tab for more information about API numbering.
OtherID	alphanumeric	0..n	Alternative identifiers for the borehole identified by the HeaderURI. The OtherID is an alternative to the APINo or WellURI.	Indicate the authority from which the OtherID is derived with a short prefix delimited by a colon ':' If identifiers from multiple different authorities are available, delimit these identifiers with the pipe character '|' This element must have a value if the APINo element is empty.	Prefixes are used to indicate the source of the OtherID. For example, if a well were known to the Arizona Geological Survey (AZGS) by the id number 1337, the prefix for 1337 would be 'azgs:' If the OtherID is a URL, the prefix would be 'http:'

*Figure 3: The **Field List** spreadsheet. The far left column corresponds with the topmost row in the Template spreadsheet; columns on the right provide descriptions for each field*

- **Data Valid Terms**: Some fields, as described in the **Field List** spreadsheet, require very specific vocabulary; data entered in these fields should generally be restricted to the terms listed on the **DataValidTerms** spreadsheet (Figure 4)

Depth reference points	
K.B.	Kelly Bushing
D.F.	Drill floor
G.L.	Ground level
S.F.	Sea floor (needs element)
Borehole shape	**terms to characterize general geometry of borehole**
vertical	default
inclined down	
curved	
horizontal	
inclined up	
unknown	

*Figure 4: A sample entry from the **Data Valid Terms** spreadsheet*

- **Reviewer Comments**: Comments from Arizona Geological Survey reviewers

Section 5: Entering Data into Content Models

To enter data, click the **Template** spreadsheet and copy your records into the spreadsheet, field by field.

For example, assume that you have information about eight well boreholes. Your information includes the following:

- An API number for each borehole
- The owner and operator of the well associated with each borehole
- The latitude/longitude coordinates for the well associated with each borehole

As indicated by the table at the bottom of this page, this information should be entered into the **WellHeader** template. Each row should constitute an individual database record, so including the top row (which contains field information), there should be nine total rows of data in the template.

Data should then be entered into the corresponding column in the **Template** spreadsheet (Table 2). Data regarding individual fields can be found on the **Field List** tab.

APINo	Operator	Mineral Owner	LeaseName	LatDegree	LongDegree
02-021-05000	Nichols, Hugh E.	State	State	33.33789906	-111.5474776
02-005-05019	Sinclair Oil & Gas	Santa Fe Pacific RR	Santa Fe	35.76820705	-112.3573439
02-001-05335	Kerr-McGee	Hortenstine-Macie	Macie Fee	35.08278252	-109.5836642
02-003-05038	Allen, Elmer R.	W. J. Davis	Davis	31.57815012	-109.7510095
02-017-05065	Texaco-Skelly-Sinclair	Navajo	Navajo	36.98885669	-110.3807293
02-021-05005	Western Oil Fields	Federal	Federal	32.94538466	-111.3689436
02-027-05011	Stewart, M. P.	Federal	Federal	32.52173669	-114.6501276
02-001-05198	Shell Oil	Navajo	Navajo	36.98311134	-109.2666708

Table 2: Sample entries from an existing template spreadsheet

It should be noted that your data may not correspond with the fields in a given content model on a one-for-one basis; if so, then a degree of schema mapping is required in order to transpose your data into the content model. See Step 2 of the data provider workflow for more information.

Section 6: Submitting Data

After you have prepared your data for submission by entering it into the template spreadsheet of a content model, you may submit your data through the AASG Geothermal Data website.

- **Login**: Click **Login** in the upper right hand corner of the AASG Geothermal Data website (Figure 5). Enter your username and password in the login window that appears. Alternatively, you may login via established electronic mail providers such as Gmail, AOL, and OpenID. To do so, simply click the appropriate icon at the bottom of the **Login** window and enter the username and password you normally use for that provider; your provider will then log you in to AASG Geothermal Data

Figure 5: Logging in to AASG Geothermal Data

- Submit your data: Mouse over the **Members** menu and click **Submit Files** (Figure 6). This will bring you a Submission page (Figure 7) on which you will provide detailed information about the data you wish to submit. Required fields are indicated by a red asterisk (*). Click the **Save** button, or click **Preview** to see what your data will look like when submitted

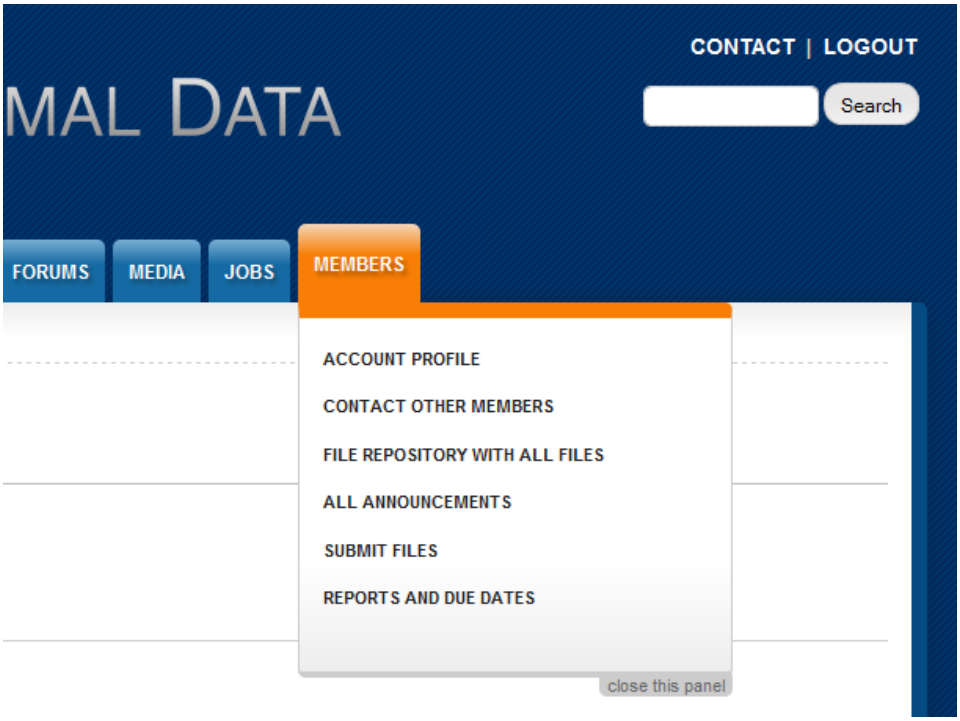

*Figure 6: The **Members** menu*

- Wait for feedback: Someone at the Arizona Geological Survey will review your data and let you know whether or not it needs to be revised. If your data needs to be revised, it will be returned to you and you may resubmit it for review

If your data is approved and does not need to be revised, the following steps need to be taken:

1. Provide the AZGS with a metadata record for your data using the USGIN Metadata Wizard. If your data is being served as a service, you will need to provide the AZGS with a metadata record for the service, as well as for the data

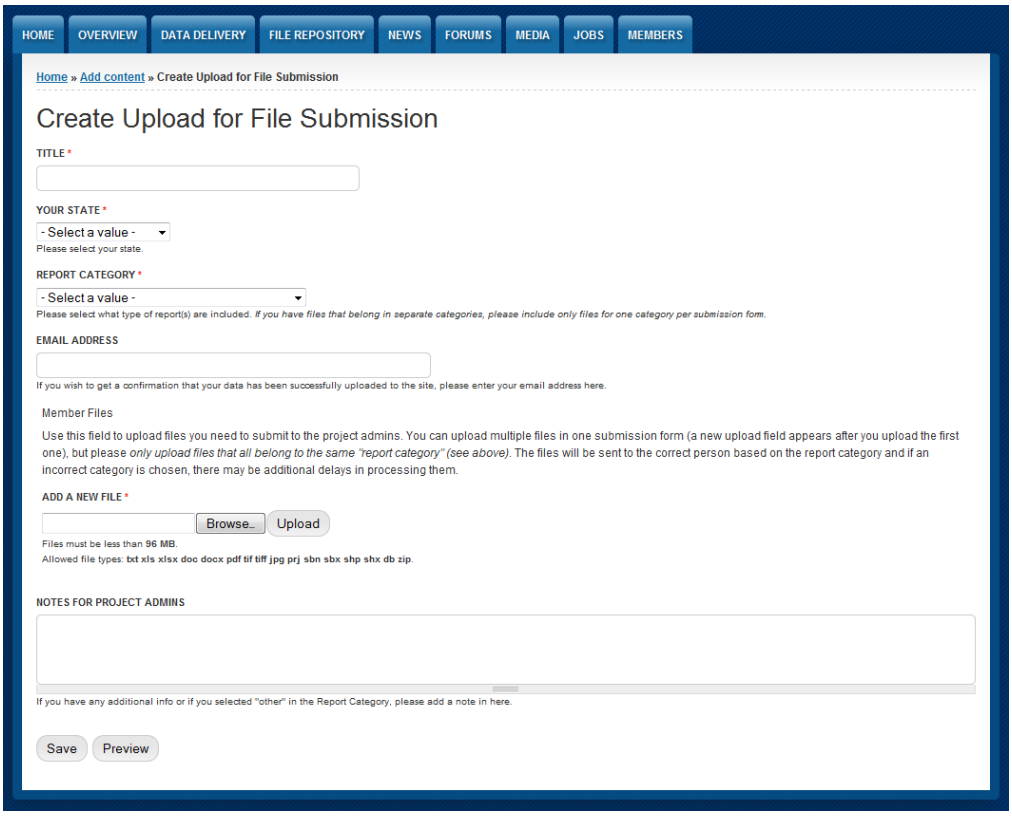

Figure 7: The Submission page

2. Get your data online. If your data is a deliverable (for example, a completed content model), it will need to be uploaded to the USGIN document repository. If your data is going to be hosted as a service, this can be accomplished in one of the following ways:

    1. If you are hosting your own data, you will need to mobilize sufficient computing resources
    2. If you are a California or Arizona subrecipient, AZGS will host your data as a service
    3. If you are a subrecipient in any other state, AZGS will notify your state of the fact that your data has been approved, and they will host your data as a service

This concludes the USGIN Content Model Tutorial.

Appendix A: Schemas, Content Models, Templates, and Interoperability

The following example will demonstrate the importance of interoperability at the data level.

Two ornithologists are recording data describing birds. Tables 3 and 4 contain their results.

Specimen ID	Species	Wingspan	Bill Length

2312	*C. Cristata*	34 cm	2.2 cm
2313	*C. Cristata*	35 cm	2.5 cm
2314	*C. Cristata*	37 cm	2.3 cm
2315	*C. Cristata*	34.5 cm	2.2 cm
2316	*C. Cristata*	36 cm	2.1 cm
2317	*C. Cristata*	36 cm	2.4 cm

Table 3: Sample data describing blue jays

Specimen_Identifier	Sp.	Wingspan	Beak_Length	Units
g:12	*Larus argentatus*	48	1.1	in
g:13	*Larus argentatus*	55	1	in
g:14	*Larus argentatus*	54	1.15	in
g:15	*Larus argentatus*	49	1.2	in
g:16	*Larus argentatus*	59	1.12	in

Table 4: Sample data describing seagulls

Both tables obviously record the same kind of data, which matches identified individuals of a given bird species with corresponding measurements of beak and wingspan.

But these tables are not interoperable because they use different schemas. In Table 3, specimens are listed under the **Specimen ID** field; in Table 4, specimens are listed under the **Specimen_Identifier** field. In Table 3, **Species** is spelled out, but the genus is abbreviated; in Table 4, **Species** is abbreviated as **Sp.** but genus is spelled out in each entry. Likewise: in Table 3, measurement units for specimen wingspan and beak measurements are provided in the same field as the measurements themselves; in Table 4, the units for these measurements are listed in a separate field. Finally: in Table 3, beak measurements are listed under the **Bill Length** field; in Table 4, they are listed under the **Beak Length** field.

All of this might not seem like much to a human operator. People can often understand and account for such discrepancies by drawing on experience and intuition (though there are occasions in which schematic differences can baffle even human operators). By contrast: unless specifically instructed to do

so, a computer cannot understand that a *beak* can also be called a *bill*, or that *species* can be abbreviated *sp.* Nor can a computer interpret units of measure unless so instructed: to a computer, there is no difference between 57 inches and 57 centimeters, unless the computer has been instructed to understand the difference.

In order to make data sharing as easy as possible for both people and computers, AASG Geothermal Data uses content models to ensure that data providers submit their data the same way. This is not to say that the content models used by the Geothermal Data project are perfect. Rather, these content models represent a practical compromise between machine readability, human readability, and the demands of the data.

To continue the example above, imagine that the Ornithology Society has decided to adopt the following schema for data describing birds:

Specimen	**Species**	**Wingspan (mm)**	**Proboscis Length (mm)**
Unique identifier for each bird measured	Genus and species of specimen; provide full names for both genus and species	Measurement of specimen from wingtip to wingtip, in millimeters (mm)	Measurement of specimen beak or bill length, in millimeters (mm)

Table 5: A content model for an ornithological schema

When fitted to the above schema, the data from Table 3 and Table 4 would appear as follows:

Specimen	**Species**	**Wingspan (mm)**	**Proboscis Length (mm)**
2312	*Cyanocitta cristata*	340	22
2313	*Cyanocitta cristata*	350	25
2314	*Cyanocitta cristata*	370	23
2315	*Cyanocitta cristata*	345	22
2316	*Cyanocitta cristata*	360	21
2317	*Cyanocitta cristata*	360	24
g:12	*Larus argentatus*	1200	27.5

g:13	*Larus argentatus*	1375	25
g:14	*Larus argentatus*	1350	28.75
g:15	*Larus argentatus*	1225	30
g:16	*Larus argentatus*	1475	28

Table 6: The data from Tables 3 and 4 fitted to the schema in Table 5

Since their data now uses the same schema, the data from Tables 3 and 4 is now interoperable and can be directly compared in a machine-friendly way, without operator interpretation.

Note that the measurements for both *C. Cristata* and *L. Argentatus* are in millimeters. This streamlines the process of searching the database, because neither the search engine nor the user need convert results when searching for bird speciments by measurement. The only downside is a certain degree of overhead on the part of data providers: in order to map their data to the schema, users must convert their measurements to millimeters.

Note also that the specimen identifiers use different formats. Though it might countermand best practice, this does not prevent interoperability either, as long as each identifier is unique.