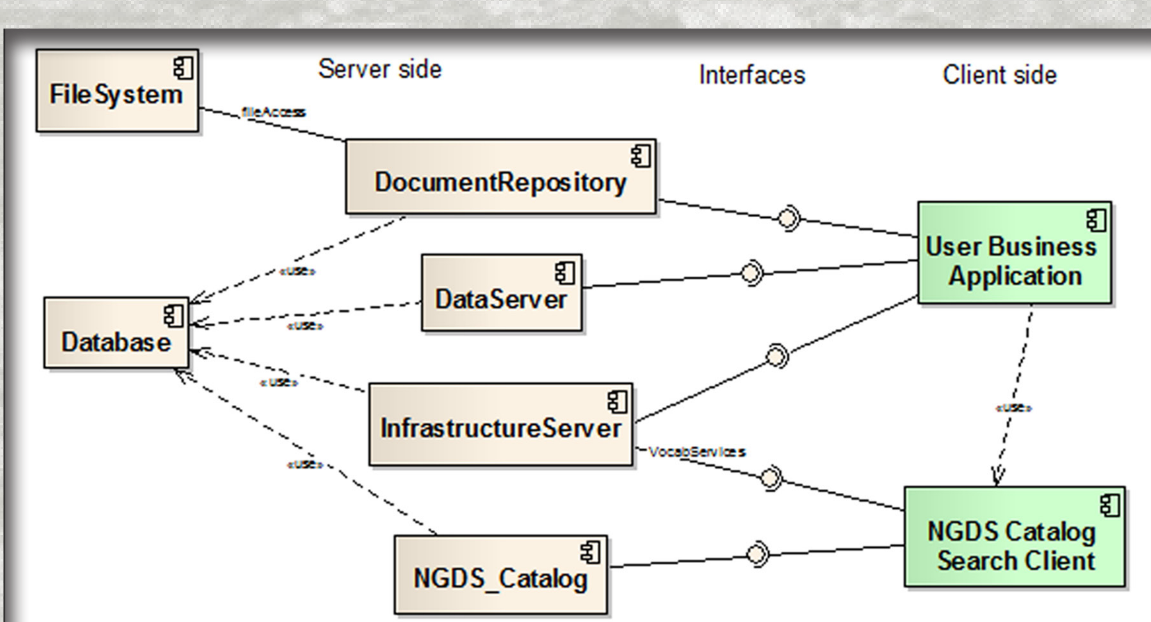
OPEN FILE REPORT OFR-10-04

Arizona Geological Survey
416 W. Congress St., Ste. 100, Tucson, AZ 85701
www.azgs.az.gov



Technical Contributions of the Geoscience Information Network for Design and Deployment of the National Geothermal Data System

Compiled by
Stephen M. Richard

2010

Technical Contributions

of the

Geoscience Information Network

For Design and Deployment of the

National Geothermal Data System

Compiled by

Stephen M. Richard

October 10, 2010

AZGS OFR-10-04

This material is based upon work supported by the Department of Energy under Award Number DE-EE0001120.

Contents

Summary

Accomplishments

The Arizona Geological Survey prepared and submitted proposals for system design, including data and metadata types, standards and protocols, and testing the system prototype, as part of a subcontract to Boise State University for development of the National Geothermal Data System (NGDS). This report is a compilation of the technical materials and recommendations developed during Phase 1 of that project for system design, data sharing and web services, data acquisition, and management and sustainability plans. This is a summary of work done between November 2009 and August 2010.

These reports and proposals document progress of an ongoing dynamic process with the goal of bringing data services online in a production information system. They include recommendations that require feedback and refinement, as well as interim reports. Other components of the data integration framework for the NGDS are in various stages of development, review, planning, or conceptualization.

AZGS met its milestones and goals for Phase 1 of the Boise State University-run NGDS project as described in the Statement of Project Objectives (SOPO Tasks 1.0, 2.0, and 4.0):

- AZGS defined specific technical deliverables and a timetable prior to the start of the project and has followed that detailed plan. In accord with subcontract provisions, AZGS submitted proposals and recommendations to project management for review and approval.

- AZGS participated in all the project meetings and webinars (Tasks 6.0 and 7.0). AZGS reviewed proposed changes to the SOPO. AZGS outlined a plan for a Technical Coordinating Council that became the basis for the Technical Working Group. The Project Management Team (Board) did not otherwise meet nor was it given any other responsibilities.

- AZGS delivered all quarterly reports and other materials as requested on time and within budget (Task 12.0)

Project Status (Specific milestones, goals, and deliverables)

Technical Activities (Tasks 1.0, 2.0, & 4.0)

After the announcement of the awarding of the project, AZGS identified the work necessary to complete each of its technical tasks in the Statement of Project Objectives (SOPO) and what AZGS deliverables would be required. This became a proposed work plan for Phase 1 that was delivered to Project Management prior to the official start date of the project. No response to the plan was received by AZGS but we proceeded to carry out the tasks and prepare the deliverables we determined were necessary to complete our requirements under the subcontract to Boise State University and the SOPO.

Since then, AZGS project members, under the technical leadership of Dr. Stephen Richard, delivered 20 documents, plans, and reviews to the Project Management, which collectively completed our goals for Phase 1. The following table summarizes the SOPO tasks, AZGS actions and deliverables, dates delivered, and their status. The results of each item are attached as an appendix to this report.

Table 1. AZGS tasks in the Statement of Project Objectives (SOPO), resultant actions and deliverables, and their status. Tasks are color coded.

SOPO Task	**SOPO Appendix 1 description of task**	**AZGS actions and deliverables**	**Date delivered**	**Status**	**Comments**
Database design (tasks 1.0 and 4.0):	Major participant in the design of the NGDS database, include data and metadata types, standards and protocols and testing the system prototype.	"Content model and rationale for minimal metadata records in a federated catalog system for geoscience resources"	3/10/2010	Waiting on action by NGDS management team	Revised 6-16-10, resubmitted
Database design (tasks 1.0 and 4.0):		List of attributes identified from heat flow data distributed to project partners	4/22/2010	Robin Penfield (GBC, UNR) responded with a list of attributes in their heat flow compilation based on SMU data, and a copy of their shapefile/spreadsheet of the compiled data. This was used to create a demonstration WFS for	

SOPO Task	SOPO Appendix 1 description of task	AZGS actions and deliverables	Date delivered	Status	Comments
				heat flow data	
Database design (tasks 1.0 and 4.0):		"Data items of interest for National Geothermal Data System." Compiled by S. M. Richard from responses by State Geological Surveys and phone conversations with NGDS Boise State project partners.	6/1/2010	Waiting on action by NGDS management team	
Database design (tasks 1.0 and 4.0):		NGDS System Architecture Diagrams	6/17/2010	Waiting on action by NGDS management team	
Database design (tasks 1.0 and 4.0):		"Draft Phase 1 report: AZGS Contribution to System Design for National Geothermal Data System," 23p.	8/13/2010	Waiting on action by NGDS management team	NGDS Director wrote that presentation of the report at the August 16, 2010 annual project meeting would be "inappropriate."
Database design (tasks 1.0 and 4.0):		outline of tasks to accomplish NGDS	6/1/2009	Waiting on action by NGDS management team	

SOPO Task	SOPO Appendix 1 description of task	AZGS actions and deliverables	Date delivered	Status	Comments
Data sharing and web services (task 2.0):	Web services: Lead role on the design, standards and protocols associated with web services, development of data catalogues for each networked database, and helping to implement web services at all networked sites. This will be done in conjunction with the core NGDS development team at Boise State, and the NGDS management team will approve all standards and protocols.	Helped develop spreadsheet for collecting input from project collaborators and collated information from the State Geological Survey spreadsheets.	November-December 2009	Distributed to NGDS project partners; input collated by Kim Kurz; follow up phone conferences. Current status unknown.	
Data sharing and web services (task 2.0):		"Metadata Wizard" distributed for review.		Waiting on action by NGDS management team	
Data sharing and web services (task 2.0):		Released "National Geothermal Data System." This document is intended to provide information about the USGIN and NGDS projects for agencies preparing proposals under US DOE DE-FOA-0000109.	7/8/2009	Waiting on action by NGDS management team	
Data sharing and web services (task 2.0):		Deployed and tested a DSpace repository as a prototype for a document repository for NGDS digital document	12/1/2009	AZGS team determined that Drupal is a better platform for NGDS purposes and has abandoned the DSpace test.	

SOPO Task	SOPO Appendix 1 description of task	AZGS actions and deliverables	Date delivered	Status	Comments
Data sharing and web services (task 2.0):	"	Deployed a GeoNetwork 2.4 catalog server application for testing, and a light weight client (CatalogConnector) for searching NGDS catalogs using the standard OCG catalog service for the web (CSW 2.0.2)	12/1/2009	Waiting on action by NGDS management team	
Data sharing and web services (task 2.0):	"	Development of a prototype document repository based on Drupal	4/1/2010	Waiting on action by NGDS management team	
Data sharing and web services (task 2.0):	"	A prototype catalog implementation with GeoNetwork went live at (http://catalog.usgin.org/geonetwork)	4/1/2010	Waiting on action by NGDS management team	
Data sharing and web services (task 2.0):	"	Developed a document describing usage of ISO metadata specifications to produce interoperable metadata, available at http://lab.usgin.org/node/295 for comment	4/1/2010	Waiting on action by NGDS management team	
Data sharing and web services (task 2.0):	"	A set of draft system architecture drawings was prepared, modeling components, nodes, and interfaces sent to BSU technical lead.	5/25/2010	Waiting on response from BSU.	
Data sharing and web services (task 2.0):	"	Formulated a plan to beta test the metadata production by registering documents at the Great Basin Center that have information for the Bradys Butte geo-	6/1/2010	More detail developed at the Technical Work Group meeting in Denver, 7-12-2010. Test metadata in prototype catalog, 8/17/2010	

SOPO Task	SOPO Appendix 1 description of task	AZGS actions and deliverables	Date delivered	Status	Comments
		thermal area.			
Data sharing and web services (task 2.0):	“	"DOE NGDS II Minimum Metadata Reporting Concepts"	6/17/2010	Superseded by USGIN metadata content recommendations document. This original document is posted on the metadata forum at geothermaldata.org. Some ideas have been incorporated into D. Morago's content model for repository documents.	
Data sharing and web services (task 2.0):	“	"Draft Phase 1 report: AZGS Contribution to System Design for National Geothermal Data System"	8/13/2010	Waiting on approval of recommendations by NGDS management team	NGDS Director wrote that presentation of the report at the August 16, 2010 annual project meeting would be "inappropriate."
Data sharing and web services (task 2.0):		Spreadsheet posted on Data Categories Working Group forum that compares the USGIN metadata recommendation with BSU proposal -http://www.geothermaldata.org/NGDSParticipants/DataCategoriesWorkinGroup/ tabid/267/ forumid/15/scope/ threads/Default.aspx	9/10/2010	Waiting comment and discussion through online forum or other contact.	

SOPO Task	SOPO Appendix 1 description of task	AZGS actions and deliverables	Date delivered	Status	Comments
Classification Systems (task 5.0):	Participate with all members in helping to implement the USGS Geothermal Resource Classification System, including possible variations on this system as needed or desired.				
Management and sustainability plans (tasks 6.0 and 7.0)	Member of the NGDS Management Board. Will participate in the design and implementation of both the management and sustainability plans for the NGDS.	Draft "Outline for an NGDS Technical Coordinating Council"	5/26/2010	Outline adopted by NGDS Director for a Technical Work Group - TechWG-but duties and membership changed	Management Board (Team) has yet to meet. No information on duties or authority of the Board (Team)
Data acquisitions (tasks 8.0, 9.0 and 10.0):	Will catalogue geothermal data from their digital and analogue files, and then digitize relevant analogue data as appropriate for the NGDS and by agreement of the NGDS Management Board. Care will be taken to not duplicate data within the NGDS.	All 45 AZGS geothermal reports scanned and posted online along with a spreadsheet of 400 Arizona geothermal technical citations (http://www.azgs.az.gov/geothermal_downloads.shtml)	late 2009	Digitization of data from files by AZGS staff is underway	
Community interaction and relationships: (task 11.0)	Will contribute to and review all materials produced for the public, including at workshops, NGDS booths at professional meetings, and web site information. Will participate in workshops and booths as appropriate and possible.	24 oral presentations; 5 newspaper, radio, and magazine interviews; 4 publications (2 in review); 1 written response (National Research Council)	December 2009 - September 2010	Deliverables met	No materials provided from BSU for review prior to release or publication.

SOPO Task	SOPO Appendix 1 description of task	AZGS actions and deliverables	Date delivered	Status	Comments
Project reporting (task 12.0)	Will contribute quarterly results and reports of activities to the NGDS Director and review the draft of the quarterly NGDS reports to be sent to DOE.	3 quarterly reports submitted	1-7-10; 4-26-10; 7-22-10	Deliverables met	No drafts or final reports received by AZGS from BSU for review.

Management and Sustainability Plans (Tasks 6.0 & 7.0)

AZGS participated in all the project meetings and webinars organized by project management. AZGS outlined a plan for a Technical Coordinating Council that became the basis for the Technical Working Group. AZGS is nominally represented on the Management Team but it has yet to meet or be charged with any responsibilities.
In March, 2010, AZGS reviewed the project “White Paper” prepared by project management.

In September, 2010, AZGS reviewed and critiqued a draft proposal from project management to repurpose the project and add significant new deliverables to AZGS.

Appendices [submission date in brackets]

A. Technical excerpt from Quarterly report, 4Q 2009 (submitted January, 2010)
B. Technical excerpt from Quarterly Report Q1, 2010 (submitted April, 2010)
C. Technical excerpt from Quarterly Report, 2Q 2010 (submitted July 2010
D. Preliminary Quarterly Report 3Q 2010
E. Brief Description of Subcontractor Roles and Responsibilities [as prepared by BSU, 10-12-09]
F. Initial Services [11-5-2009]
G. Summary of NGDS Surveys [1-27-2010]
H. NGDS Data Acquisition Plan [2-10-2010]
I. "Some thoughts" memo [3-2-2010]
J. Content model and rationale for minimal metadata records in a federated catalog system for geoscience resources [3-10-2010]
K. Data items of interest for National Geothermal Data System [4-17-2010]
L. Summary, NGDS components and interfaces [4-20-2010]
M. Component deployment in NGDS client and server nodes NGDS [4-20-2010]
N. NGDS-DOE Minimum Metadata Reporting Requirements [4-20-2010]
O. AASG-NGDS Statement of Work outline [4-20-2010]
P. Top level NGDS components [4-20-2010]
Q. Heat flow data demonstration plan [4-22-2010]
R. Outline for an NGDS Technical Coordinating Council [5-26-2010]
S. Draft Phase 1 report: AZGS Contribution to System Design for National Geothermal Data System [8-21-2010]
T. Comments on "NGDB Repository Metadata for Well Logs" [9-15-2010]
U. Metadata Crosswalk: BSU Repository – USGIN Recommendations

Appendix A. Technical excerpt from Quarterly report, 4Q 2009 (submitted January, 2010)

Stephen Richard helped Kim Kurz at BSU develop spreadsheet for collecting input from project collaborators. Kim distributed to collaborators, and collated results. The information was a start for indicating the kinds of services required, but more specific info is needed. Steve Richard started collating information from the spreadsheets AZGS collected from State collaborators for NGDS 2 to get more specifics on the kinds of information resources we will need to accommodate.

Ryan Clark has deployed a DSpace repository as a prototype for a document repository for NGDS digital documents. We are currently testing this prototype by loading some AZGS documents.

Wolfgang Grunberg and Stephen Richard have been working on an ISO19115/119/139 metadata profile document to specify metadata content and encoding for interoperable catalog services. We have deployed a GeoNetwork 2.4 catalog server application for testing, and a light weight client (CatalogConnector) for searching NGDS catalogs using the standard OCG catalog service for the web (CSW 2.0.2).

AZGS has acquired a new Windows 2008 server which Ryan Clark has set up and is configuring for web services and web hosting.

We anticipate that the catalog system will be deployed during January 2010, with development work changing focus to metadata entry tools, metadata acquisition for document holdings, and identification of some key data services to work on developing.

Lee Allison testified to the Arizona Legislature about digital databases using the NGDS project capabilities as the demonstration mode.

Lee Allison spoke and met with representatives from Alta Rock Energy about requirements for them to share their data with the NGDS.

Appendix B. Technical excerpt from Quarterly Report Q1, 2010 (submitted April, 2010)

Major development focus during this quarter has been on the catalog system and document-management system. The catalog system development has been pursued on two fronts—a catalog service server implementation, and profiles for metadata content and encoding. The document management system work included evaluation of options, testing of DSpace and Drupal, and development of a prototype document repository based on Drupal.

Under the auspices of the NSF- funded INTEROP - GIN project, we have already done research and testing of various applications that implement the OGC CSW service, and have focused our attention on Geonetwork opensource (http://geonetwork-opensource.org/) as the implementation of choice for developing a catalog prototype. A prototype catalog implementation with GeoNetwork is now live at (http://catalog.usgin.org/geonetwork). Geonetwork can accommodate various metadata formats (FGDS, Dublin Core/csw:record, ISO 19139, ISO on ebRIM), and can harvest metadata from various sources including CSW catalogs, Geonetwork catalogs, and from OGC web service getCapabilities documents. There are still plenty of hiccups in the system but based on our investigations, Geonetwork provides the most robust, functional, and extensible environment for a community catalog implementation that the NGDS can provide to participating agencies and interested collaborators who want to implement a catalog server to publish metadata to the system.

We have developed a document describing usage of ISO metadata specifications to produce interoperable metadata. The current version of this document is available at http://lab.usgin.org/node/295. We are seeking community input to improve this document, and anticipate that the revised version will be synchronized with the ISO metadata profile in development with the Energistics Energy Industry Metadata Working group (http://www.energistics.org/metadata-work-group), an activity that we are participating in under the auspices of this project. In addition, a simplified description of the minimum content for NGDS metadata is under review by the NGDS developers.

We participated in several phone conferences and review of a questionnaire we distributed to project partners to analyze data requirements. In addition, we carried out a survey of the data proposed for submission to NGDS by state geological surveys. Based on this user input (see attached DataItemSummary.doc), we are planning the architecture for types of services and different kinds of content expected to populate NGDS.

One of the immediately available resources that nearly all project partners have ready to start integrating is online documents. We are planning to make these part of the system by generating metadata descriptions of each document according to an NGDS metadata content model so that the metadata are interoperable. Metadata will be published by making them available for search or harvest on a public server. A requirement of the metadata will be to include a URL that allows retrieval of a described document via the Web. By aggregating metadata describing documents provided by various partners that can be searched by a single client, users will be able to discover and obtain documents from any NGDS compliant document repository.

For partners that do not currently have an implemented repository, we have researched various options, and have selected a simple repository implementation using Drupal, a free, open source content management system. A prototype repository is currently being tested at AZGS.

Appendix C. Technical excerpt from Quarterly Report, 2Q 2010 (submitted July 2010)

Development focus in this quarter has been on the document repository, simplifying production of metadata, discussion of use cases to develop recommendations for phase one report, and metadata profile development.

The Drupal document repository implementation was demonstrated to the Boise State Team at the beginning of the quarter (4/21). Based on feedback and discussion, development and debugging continued into the quarter. It was recognized that a separate, standalone version of the application would be useful for generating metadata for resources that do not have any existing metadata, and are already in an online repository. A prototype for this tool has been developed using the metadata production components from the document repository, and is currently able to export the metadata content in ISO19139 or FGDC XML; we are referring to this as the Metadata Wizard. The intention is that other metadata encodings (DIF, EML, csw:Record…) can be supported as well.

The data item summary was distributed at end of last quarter to NGDS project partners, but so far we have gotten no feedback. This compilation of data items is meant to be starting point for identifying the features or entities that will be the principal elements in interchange documents used to move data around the system (not a data model for any particular database).

A phone conference was held with the BSU team on 4/8/2010 to discuss development activities, mostly on the document repository approach and metadata.

On April 22 a list of attributes identified from heat flow data that we were able to locate was distributed to project partners with a request for comment and input as to whether these provided the necessary information for heat flow measurement. Robin Penfield (GBC, UNR) responded with a list of attributes in their heat flow compilation based on SMU data, and a copy of their shapefile/spreadsheet of the compiled data. This was used to create a demonstration WFS for heat flow data, based on phone conference plan from 4/21.

In a phone conference with the BSU Team (plus David Cuyler, Sandia) in late June, we formulated a plan to beta test the metadata production by registering documents at the Great Basin Center that have information for the Bradys Butte geothermal area. This plan was filled out in more detail at the Technical Work Group meeting in Denver (07/12/2010), and will be a third quarter activity. The intention is to test the proposed metadata content model distributed for comment in the first quarter with a collection of real information resources to determine if sufficient information can be captured, and to test the Metadata Wizard user interface.

Development has continued on the Energistics Energy Industry Metadata profile, with the intention of aligning a NGDS metadata profile with the Energy Industry profile. This will be an ISO 19115/19119/19139 profile similar to the USGIN profile. The schedule is to have a draft profile document in December.

A set of draft system architecture drawings was prepared, modeling components, nodes, and interfaces. These drawings were sent to Christian Loepp on 5/25/10 for comment, and distributed to the technical teams of NGDS-related projects at BSU, SMU, USGS after the 6/16/10 phone conference with the intention of stimulating discussion of the system architecture and production of formal architecture design diagrams for the Phase one report. We are awaiting feedback and discussion.

Minor updating and debugging of the demonstration Geonetwork instance was required during the quarter. Wolfgang Grunberg worked with Jordan Hastings of the Great Basin Center to develop a workflow for importing FGDC metadata into the catalog by using XSLT transformation to ISO19139 (using the NOAA developed transformation code as a starting point). The transformation was successfully implemented and tested, but it was determined that the actual content of the GBC metadata needed to be updated, so the project was put on hold for GBC to determine how to update their metadata.

Appendix D. Preliminary Quarterly Report 3Q 2010

The third quarter project report is in preparation and will be delivered within the stipulated delivery date. The following are some key actions that occurred during the quarter.

Lee Allison and Steve Richard met with project managers in Salt Lake City on August 15 prior to the projects annual meeting that began the following day. The goal was to clarify our duties in light of the changes in the project. Agreement was reached on using the Financial Risk Assessment tool as a demonstration test bed for system design and implementation.

Allison and Richard presented overviews of the GIN concept and summaries of deliverables at the projects annual meeting in Salt Lake City, on August 16-17, 2010.

AZGS reviewed, critiqued, and responded to proposed significant revisions of the Statement of Project Objectives (SOPO).

On September 10, we posted a spreadsheet to the project Data Categories Working Group forum (Data Categories http://www.geothermaldata.org/NGDSParticipants/DataCategoriesWorkingGroup/tabid/267/forumid/15/scope/threads/Default.aspx) that compares the USGIN metadata recommendation with that presented by Kim Kurz at the project meeting in Salt Lake City in August.

On September 14, we reported that at the recent IUGS CGI Interoperability Workgroup meeting, the group started a new task to develop xml schema for flat file (simple feature) gml views of GeoSciML to facilitate use with layer-based GIS clients (like ArcGIS). Steve Richard supported this idea because it corresponds nicely with the kinds of simple feature services we have been talking about for the NGDS data services. Some web pages have been started on the GeoSciML twiki for discussion

(https://www.seegrid.csiro.au/twiki/bin/view/CGIModel/GeoSciMLThematicView),

and we distributed a spreadsheet with the current proposal. We requested contributions and discussion on the GeoSciML twiki in hopes that the CGI schema can be used directly for NGDS geologic unit and structure services.

On September 15, we provided detailed comments on the draft “NGDB Repository Metadata for Well Logs” document circulated by Daniella Morgos at BSU.

Appendix E. Brief Description of Subcontractor Roles and Responsibilities [as prepared by BSU, 10-12-09]

The Geothermal Data Consortium (GDC) is a partnership that combines critical geoinformatics and geothermal experience and expertise. This partnership reflects a fundamental understanding of critical success factors and institutional barriers associated with creation and operation of the NGDS. The GDC recognizes that building the NGDS will not be a simple task but the members have made the commitment to work together toward a common goal for the development, maintenance, and growth of a true distributed "data system". The GDC expertise and consequent general roles and responsibilities are listed below.

Geoscience Information Network (GIN), Arizona Geological Survey:

Database design (tasks 1.0 and 4.0):

Major participant in the design of the NGDS database, include data and metadata types, standards and protocols and testing the system prototype.

Data sharing and web services (task 2.0):

Web services: Lead role on the design, standards and protocols associated with web services, development of data catalogues for each networked database, and helping to implement web services at all networked sites. This will be done in conjunction with the core NGDS development team at Boise State, and the NGDS management team will approve all standards and protocols.

Primary responsibility to work with state geological surveys to provide easy access and interoperability to survey-hosted data

Classification Systems (task 5.0):

Participate with all members in helping to implement the USGS Geothermal Resource Classification System, including possible variations on this system as needed or desired.

Management and sustainability plans (tasks 6.0 and 7.0)

Member of the NGDS Management Board. Will participate in the design and implementation of both the management and sustainability plans for the NGDS.

Data acquisitions (tasks 8.0, 9.0 and 10.0):

Will catalogue geothermal data from their digital and analogue files, and then digitize relevant analogue data as appropriate for the NGDS and by agreement of the NGDS Management Board. Care will be taken to not duplicate data within the NGDS.

Community interaction and relationships: (task 11.0)

Will contribute to and review all materials produced for the public, including at workshops, NGDS booths at professional meetings, and web site information. Will participate in workshops and booths as appropriate and possible.

Project reporting (task 12.0)

Will contribute quarterly results and reports of activities to the NGDS Director and review the draft of the quarterly NGDS reports to be sent to DOE.

Appendix F. Initial Services [11-5-2009]

- Document discovery service (DDS): search for documents that are in all the repositories that are in all the collaborators here: things that exist that are out there in spreadsheets, get metadata describing what they are and get acquiring what they are (7)
- Online Map discovery services (OMDS): a lot of people say they have some maps, scan, georeference; similar to the document discovery service, find an online version of a paper map to put in your GIS project (8)
    - Temperature gradient visualization leads to a
    - heat flow map
- Feature service for leases (0)
- Borehole (well) locations: filter boreholes for the purpose of the drilling, information obtained from the boreholes (3)
- Borehole bottom-hole temperature, bottom-hole location (x,y,z) (1)
- Temperature gradient index to boreholes (0)
- Geothermal Springs (0)
    - Chemistry
    - Flow
    - Salinity
- If you want to get all the data for some area that I can from all the different sources without having to go out to each individually (theses, some from multiple states, etc) (points 1 and 2; 2)
    - Sit at computer and pull information from as many different servers as possible, some reports have different information (ex: chemistry on area with different elements) (0)
    - Desktop is meant as initial portal
- ****Use case: report data, what does the system expect from the reporter**** format, protocol, and style for how we want the data delivered (data acquisition phase 1) (everyone)
    - Have a home for this stuff and tell people how to give it to us
    - To what precision do we ask for data
- Progress map: contributions to geothermal dataset, derived from existing, where we have datasets, maps, multiple layers; interesting cartographic problems (1=100)

Appendix G. Summary of NGDS Surveys [1-27-2010]

DataItem	SumOfCount	Needs clarification	services	Needs modeling
Metadata	1174137		CSW, ISO19139	
Borehole temperature data	571355	clarify what kinds of measurements	WFS: observations, temperature; WFS: WITSML log	temperature observations may need some clarification
Borehole lithology log	94802		WFS: mapped feature, borehole with GeoSciML geoUnit desc; WITSML log?	
Water source characterization	84812	what are characteristics	CUAHSI soft-typed data approach? WFS:observation, chemical property	water source feature
Document	15538		for documents, catalog service using location; ISO19139 metadata	
Developed geothermal system feature	13654	what are characteristics	for documents, catalog service using location; ISO19139 metadata: WFS geothermal system feature	developed geothermal system feature
Water Chemistry	7259		CUAHSI soft-typed data approach? WFS:observation, chemical property; EarthChem chemistry xml?	
Bottom hole temperature	5952		WFS: observations, temperature	
Heat flow measurement	2724		WFS: observations, heat flow	
Digital well log	1388		WFS: WITSML log; OGC coverage	
Rock chemistry	989		WFS:observation, chemical property; EarthChem chemistry xml?	
Hot spring description	649	what are characteristics	for documents, catalog service using location; ISO19139 metadata; WFS: hot spring xml?	
Drill stem test	400	what are characteristics	WFS: observation???	
Text description	211	what are units of delivery	are these documents or text to put in a database?	
Geologic Unit geothermal characterization	150		WFS: sample, GeologicUnit: physical properties	geothermal rock characterization
Volcanic vent description	138		WFS: geologicFeature: vent	volcanic vent feature
Thermal conductivity measurement	100		WFS: sample, GeologicUnit: physical properties	
Active Fault	97		WFS: GeoSciML fault	
Flow rate	92		WFS: observation: flow rate	
Geologic map	45		WMS for scans, WFS for vector data; File discovery from catalog and download from	

			repository for file-based (shapefiles, MIF, e00, fileGeodatabase, GML)	
Permeability	18		WFS: sample, GeologicUnit: physical properties	
Enhanced geothermal system feature	17	what are characteristics	WFS: Geothermal feature	Enhanced geothermal system feature
Resource suitability map	12		WMS for scans, WFS for vector data; File discovery from catalog and download from repository for file-based (shapefiles, MIF, e00, fileGeodatabase, GML)	resource suitability feature
Alteration description	5		WFS: sample, GeologicUnit: GeoSciML	
Fluid inclusion data	5		WFS: observation: equilibration temp, etc	FLINC observation results data type
Geothermal map	2	what does this show	WMS for scans, WFS for vector data; File discovery from catalog and download from repository for file-based (shapefiles, MIF, e00, fileGeodatabase, GML)	
Aquifer temperature map	1		WMS for scans, WCS--coverage, temperature/depth grid	
Crustal Stress data			WFS: observation, stress state tensor	stress measurement
Gravity data			WFS: observation, gravity station; WCS gravity grid (isostatic, bouger, …); WMS for scanned map	
Intrusive body with heat			WFS: Mapped feature, intrusive body, geolgoic unit description	
Earthquake epicenter			WFS: mapped feature: 3D point, EQ epicenter feature	earthquake epicenter feature
Samples			WFS: sample; use SWE xml	
Production statistics record		what are characteristics	energistics ProdML (production xml for oil and gas…)? Any CUAHSI stuff on water well production?	geothermal production record

Appendix H. NGDS Data Acquisition Plan [2-10-2010]

The National Geothermal Data System must provide online resources to make it easy for users to extract assess and synthesize data according to criteria they select. The framework for implementing this requirement is a service-oriented architecture, with a community of data providers exposing information through standardized internet-accessible interfaces, a community of software developers building applications that will utilize the information resources available to the community, and a community of users taking advantage of the software and information to develop geothermal resources. A key component in this system will be the catalog services through which data providers register the availability of resources, and users discover, evaluate, and access resources.

This data acquisition plan is a road map for building the information infrastructure that is the foundation of this system. In a nutshell, the steps in this plan are:

1. Identify the kinds of information to be made available through the system.

2. Prioritize acquisition according to availability, importance for geothermal resource evaluation and development, and difficulty of acquisition.

For the various kinds of information requirements, the development team will need to specify the types of data and metadata that need to be accommodated, their variables, units of measure, and required controlled vocabularies. Some important requirements include:

1. Include all known or potentially important base-level data types (and metadata) in the system;
2. Ensure interoperability among data sets with members adopting common standards and protocols. Interoperability means in practice that software will use the same access protocol for a given kind of information from any NGDS data provider, without any provider specific customization.
3. The data schema must be vetted with stakeholders and expanded as necessary
4. Establish criteria that can be used to filter data and categorize it according to established and user-defined quality levels. These quality filters will vary depending on the type of data and their targeted use.

The process of identifying of kinds of information to be made available will be pursued on two fronts. First, we will poll the consortium member to get an inventory of the resources that they have, and how those resources are currently distributed. This polling will be done through a questionnaire and through verbal interviews of information managers with the various organizations. Second, as the development of the geothermal desktop application will be a phased process, with various use cases specified for early development. The data acquisition process will be planned to focus on delivering information to enable use cases being implemented by the desktop in order to make utilization of the desktop application a reality as soon as possible.

The development team (Boise State and AzGS) will compile the results of polling to produce a list of information resources. These will be analyzed to categorize them according to the kinds of web services required to deliver the information.

The actual data acquisition process will involve several steps. First an application profile for the service or services to deliver a particular kind of data will have to be developed. In most cases, we anticipate that existing standard services like the OpenGeospatial Consortium (OGC) Web Feature, Map, or Coverage services (WFS, WMS, WCS), will provide the necessary framework for services we require. The NGDS development group will need to develop profile documents specifying the details of how a particular type of information (e.g. borehole temperatures, water chemistry analytical data) will be encoded. Once a profile is in place for a particular data resource, the next step is working with the data-providing organization to implement the service with their data.

The actual mechanics of bringing particular datasets on line will be dependent of the format of existing data, and the IT resources of the data owner. Some organizations may choose to implement web services on their own servers to expose datasets

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

Registration of a dataset in the catalog, and its availability online will constitute 'data acquisition'. Thus, implementation of the catalog as an operational service will need to be one of the first steps in system implementation.

Appendix I. “Some thoughts” memo [3-2-2010]

In response to e-mail from Walt Snyder (3/1/2010 10:09 PM) with attached draft white paper outlining vision for NGDS.

From: Stephen M Richard <steve.richard@azgs.az.gov>

Date: Tue, 02 Mar 2010 12:07:36 -0700

To: Walter Snyder <wsnyder@boisestate.edu>, jmoore@egi.utah.edu, Lee Allison <lee.allison@azgs.az.gov>

Walt, Joe—

I thought since Walt and Joe are both authors on the white paper I should respond to both. I can't find Phillip Bandy or I'd copy him as well--can you introduce me? Anyway, the white paper is a good idea and touches on a lot of the important points. I've attached a copy with some pdf notes inserted with comments.

In particular the section on web services seems to indicate some significant divergence in understanding of the system architecture. To frame that discussion in a useful way, there first needs to be some agreement on what the NGDS is. Here's my perception:

"The National Geothermal Data System (NGDS) is an information system that consists of a linked collection of geothermal data providers, data archives, and software applications using shared standards and protocols for information interchange, with a management framework to maintain system integrity and longevity. The purpose of this system is to provide Web-based, open access to data and information pertinent to increasing geothermal's contribution to the national energy portfolio."

Given that, here's my rewrite of the web services section:

Web services are the mechanism for data publication and circulation in the system. Partners will implement sites that offer NGDS-web services, including data and catalog services. Participating databases and affiliates will make their information resources and capabilities known by publishing metadata describing the resources through the catalog service. They may also publish data resources through web services by arrangement with partners. NGDS client applications will consume NGDS conformant web services developed by any data sites. These services will allow users to locate, evaluate, and acquire targeted data from the NGDS. The NGDS-core will maintain and expose a catalog service that will be the authoritative registry of NGDS services, data, and other resources. This registry will be maintained by harvesting metadata from other catalog services, or by working with data providers to load existing metadata or create metadata for undocumented resources.

Metadata from NGDS catalog will identify an authoritative ‘host of record’ for NGDS information resources. The NGDS-core will provide a resource for hosting authoritative services if the data originators choose not to. Data or metadata harvested, cached, and possibly served by other nodes in the system must identify the data originator as well as the NGDS authoritative host for the information. NGDS-core will also provide a registry and repository for normative service specification documents. The geothermal desktop application will enable users to utilize NGDS services to access data and information pertinent to increasing geothermal contribution to the national energy portfolio.

The web services are fundamental to data discovery and access by users. Data are not published (part of the system) unless they are registered in the catalog and accessible through a public interface (web service or web-accessible API).

The web services are orthogonal to questions of archival data management or data acquisition.
An *essential* part of the system architecture is the decoupling of client applications and data archives from data provider systems through the use of interfaces. This keeps the components portable and reusable, keys for longevity.

Steve

Appendix J. Content model and rationale for minimal metadata records in a federated catalog system for geoscience resources [3-10-2010]

Subsequently revised, current version is available at http://lab.usgin.org/sites/default/files/profile/file/u4/USGIN_MetadataRecommendationsGeoscienceResources_v1.03.pdf. Title changed to “Metadata recommendations for Geoscience Resources”.

USGIN.ORG US GEOSCIENCE INFORMATION NETWORK

USGIN
U.S. Geoscience Information Network

Content model and rationale for minimal metadata records in a federated catalog system for geoscience resources

Title:

Content model and rationale for minimal metadata records in a federated catalog system for geoscience resources

Latest released version:

tba

Creator:

USGIN Standards and Protocols Drafting Team and NGDS developer team

Editors:

Stephen M. Richard and …… whoever edits

Creation date:

2010-03-10

Document Status:

Draft V. 0.1.1

Publisher:

Arizona Geological Survey, Boise State University

Description:

This document describes requirements for a minimal ISO19139 metadata record to describe resources of interest in an interoperable geoscience information system, including dataset, services, physical samples, maps and other publications, and software applications. This document is meant to provide guidance for metadata content requirement and their relationship to metadata use cases.

Contributor:

See acknowledgements

Document Identifier:

tba

Notices

Neither the USGIN project, nor any of the participating agencies, take any position regarding the validity or scope of any intellectual property or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; neither do they represent that there has been any effort to identify any such rights.

This document and the information contained herein is provided on an "AS IS" basis and USGIN DISCLAIMS ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Revision History

Version	Date	Comments	By
0.1	2010-03-10	Initial draft document	Stephen Richard

Acknowledgement

Many individuals and organizations have contributed to or inspired the development of these Guidelines.

The USGIN Standards and Protocol Drafting Team include (alphabetically):

Ryan Clark - AZGS
Wolfgang Grunberg – AZGS
Dominic Owen - AZGS
Stephen M. Richard – AZGS

The NGDS Development Team include (alphabetically):

Kim Kurz – BSU
Christian Loepp -- BSU
Charlla Adams – BSU
Daniella Morgos -- BSU
Walt Synder -- BSU

Funding Provided by (chronologically):

National Science Foundation under EAR-0753154 to the Arizona Geological Survey acting on behalf of the Association of American State Geologists; 2009.
US Department of Energy under award DE-EE0001120 to Boise State University; 2010.

Contact Information

Arizona Geological Survey
416 W. Congress St., Suite 100.
Tucson, Arizona 85701-1381
Phone: 520.770.3500
Fax: 520.770.3505

Email

metadata@usgin.org

Online

http://usgin.org
http://lab.usgin.org

Table of Contents

Tables and Figures

Introduction

A key component of a distributed information network is a catalog system, a collection of resources that allow data and service providers to register resources, and data consumers to locate and use those resources. Currently, many online catalogs are web pages with collections of URLs for services, or services are discovered accidently or by word of mouth. The vision is to enable a web client (portal) to search across one or more metadata registries without having to configure the client individually for each of the registries that will be searched. Thus, metadata providers can focus on data development, without having to also develop web clients to enable search of that metadata.

Production of quality metadata is time consuming, tedious, and gets little recognition, but good metadata is an important component to build a useful federated information system. Existing metadata standards are large complex information schema designed to account for any kind of resource description someone might want to create. This complexity makes them hard to use. Our goal is to define a minimum content requirement that can be described in relatively simple language with common sense explanation of what the purpose of the content is. The scoping of the requirements is based on a collection of scenarios for how the metadata is intended to be used.

This document then presents recommendations for standardized encoding of that content using existing metadata standards from the International Organization for Standardization (ISO), the US Federal Geographic Data Committee, and the OGC Catalog Service for the Web (CSW) using a schema for Dublin Core metadata. The approach to information encoding is based on the USGIN guidelines for encoding metadata using the ISO 19139 xml schema implementation of ISO19115/19119.

Normative References

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

USGIN_ISO_Metadata_1.1.3 USGIN profile of ISO content models (ISO19115 and ISO19119) and encoding (ISO19139). Access at http://lab.usgin.org/node/235.

ISO 19115 designates these two normative references:

- ISO 19115:2005, *Geographic information - Metadata*
- ISO 19115/Cor.1:2006, *Geographic information – Metadata, Technical Corrigendum*

ISO 19119 designates these normative references:

- ISO 19119:2005, *Geographic information - Services*
- ISO 19119:2005/Amd 1:2008, *Extensions of the service metadata model ISO 19108 designates*:
- ISO 19108:2005, *Geographic information – Temporal Schema*

ISO/TS 19139:2007, Geographic information - Metadata – XML Schema Implementation

ISO 10646-1, Information technology — Universal Multiple-Octet Coded Character Set (UCS) — Part 1: Architecture and Basic Multilingual Plane

RFC 2119, Key words for use in RFCs to Indicate Requirement Levels, Network Working Group, 1997.

Purpose

This document is intended to provide guidance on the minimum metadata content required to meet the use requirements for USGIN metadata. The intention is to reduce the daunting complexity of the ISO metadata specifications to a manageable level to promote development of interoperable metadata records for a federated resource catalog system.

Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in Internet RFC 2119.

Application profile: a schema that consists of data elements drawn from one or more namespaces, combined together by implementers, and optimized for a particular local application. (Rachel Heery and Manjula Patel, 2000, http://www.ariadne.ac.uk/issue25/app-profiles/)

Catalog application: Software that implements a searchable metadata registry. The application must support the ability to register information resources, to search the registered metadata, to support the discovery and binding to registered information resources within an information community.

Codelist (also as Code list): a controlled vocabulary that is used to populate values for an xml element.

Data product specification: a definition of the data schema and value domains for a dataset. The data schema specifies entities (features), properties associated with each entity, the data type used to specify property values, cardinality for property values, and if applicable, other logical constraints that determine data validity. Value domains are specified for simple data types—strings or numbers, and may include controlled vocabularies for terminology required to specify some properties.

Dataset series: collection of datasets sharing the same product specification (ISO 19115). ISO 19115 does not define product specification. For the purposes of USGIN, a product specification defines a data schema, any required controlled vocabularies, and recommended practices for use of schema (see Data product specification).

Dataset: an identifiable collection of data (ISO19115). USGIN refines this concept to represent a collection of data items in which individual data items are identified and accessible. USGIN extends the concept of data items to include physical artifacts like books, printed maps and diagrams, photographs, and material samples--any identifiable resource of interest. DCMI definition is "Data encoded in a defined structure" with additional comment "Examples include lists, tables, and databases. A dataset may be useful for direct machine processing." Metadata for the collection is a different type than metadata for individual items in the collection (dataset vs. features). Criteria for what unifies the collection are variable (topic, area, author...). Data items may represent intellectual content -- information content and organization (data schema) -- or may represent particular manifestations (formats) of an intellectual artifact.

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

Metadata element: a discrete unit of metadata (ISO 19115), an attribute of a metadata entity. A metadata element contains some content specifying the value of the element; this content may be simple—a number or string, or may be another metadata entity.

Metadata entity: a named set of metadata elements describing some aspect of a resource.

Metadata register: an information store that contains a collection of registered metadata records, maintained by a metadata registry. (ISO 11179)

Metadata registry: an information system for assignment of unambiguous identifiers to administered metadata records. (ISO 11179)

Metadata section: Part of a metadata document consisting of a collection of related metadata entities and metadata elements (ISO 191115).

Metadata: data about a resource in some context. Generalize from ISO 11179 definition of metadata, which constrains the scope to data about data. For USGIN purposes, metadata may describe any resource—including electronic, intellectual, and physical artifacts. Metadata represent resource characteristics that can be queried and presented for evaluation and further processing by both humans and software.

Profile: set of one or more base standards and - where applicable - the identification of chosen clauses, classes, subsets, options and parameters of those base standards that are necessary for accomplishing a particular function [ISO 19101, ISO 19106]

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

Service metadata: metadata describing the operations and information available from a server.

Source Specification: The specification or standard that is being profiled.

User Community: A group of users, e.g. within a supply-chain industry, the members of which decide to make a similar usage of the source specification in order to be able to interoperate.

Note that throughout this document, the names of xml elements are shown in `this typecase`. Long X-paths have been broken with non-breaking hyphen characters. Note that hyphens are not used in any xml attribute or element name, so if they appear in the text, they are strictly for better text wrapping. In Xpath expressions `/../` indicates that some elements have been omitted from the path.

ISO Schemas Location

ISO I9139 xml schemas are in an online repository at http://schemas.opengis.net/iso/19139/. Two versions are posted: 20060504 and 20070417. Unfortunately, these two directories both contain schema with the same target namespace, so there is no clear way to distinguish applications that are based on one or the other. The medatadaEntity.xsd in the two directories is identical; other schema have not been compared (but see discussion paper gin2009-005 at http://lab.usgin.org/node/269). The 20070417 directory contains schema implementing ISO Technical Specification 19139:2007 (dated 2007 Apr 17), which appear to include the changes from ISO 19115:2003 Cor 1;2006(E), but this is not declared in any included documentation (need metadata on the metadata schema!).

The 20070417 version of the ISO 19139 schemas references GML 3.2.1. However, there is no mention of the SRV namespace (http://www.isotc211.org/2005/srv) anywhere in this ISO 19139 version. The SRV namespace is where, in our metadata documents using the 2006 version, we specified all our information about dynamic, online services such as WFS and WMS, so the 20070417 version is not useful for metadata catalogs that register services.

In order to create metadata for both static datasets and dynamic, online services and for use with CSW, the OGC created an xml schema that merges the schema for ISO19115 (dataset metadata) and ISO19119 (service metadata) (see section D.1.5, page 105 in OGC 07-045). The way that was accomplished was by creating a schema located at http://schemas.opengis.net/csw/2.0.2/profiles/apiso/1.0.0/apiso.xsd. This schema simply imports .. iso/19139/20060504/gmd/gmd.xsd and .. iso/19139/20060504/srv/srv.xsd. Thus for CSW 2.0.2 implementations, the 20060504 versions of the ISO19139 schema must be used.

Abbreviations

CSW	Metadata Catalog for the Web. Also abbreviated as CS-W and CS/W
GeoSciML	Geoscience Markup Language
GML	Geographic Markup Language
GUID	Global Unique Identifier
IEC	International Electrotechnical Commission
ISO	International Organization for Standardization
UML	Unified Modeling Language
URI	Universal Resource Identifier

USGIN	U.S. Geoscience Information Network
WCS	Web coverage Service
WFS	Web Feature Service
XML	eXtensible Markup Language

Use cases, scenarios, requirements

This section includes a number of user scenarios for how we intend USGIN metadata to be used, and discussion of several basic approach requirements that guide determination of minimum requirements. At its heart, the problem is to find resources of interest via the internet, based on criteria of topic, place, or time, evaluate resources for an intended purpose, and learn how to access those resources. Detailed metadata describing a resource data schema, describing service or application operation, or providing detailed descriptions of analytical techniques and parameter are outside the scope intended for USGIN metadata. Our contention is that this more domain/resource specific type information is better accounted for with linked documents utilizing schema appropriate to those specific resource. Some examples include OGC getCapabilities, WSDL, and ISO19110 feature catalogs.

A user specifies a geographic bounding box or one or more text keywords to constrain the resources of interest, and searches a metadata catalog using these criteria. The user is presented with a web page containing a list of resources that meet the criteria, with links for each resource that provide additional detailed metadata, and direct access to the resource if an online version is accessible, e.g. as a web page, Adobe Acrobat document, or online application.

A client application provides user with a map window that contains some simple base map information (political boundaries, major roads and rivers). User wishes to assemble a variety of other data layers for a particular area for some analysis or data exploration, e.g. slope steepness, geologic units, bedding orientation, and vegetation type for a hazard assessment. User centers map view on area of interest, then using an 'add data' tab, accesses a catalog application that allows them to search for web services that provide the desired datasets. After obtaining the results and reviewing the metadata for the located services, user selects one or more to add to the table of contents for the client application. Response from catalog has sufficient information to enable the client application to load and use the resource (e.g. serviceType, OnlineResourceLinkage). More concrete instances of this case would be finding Web Map services to add as layers in an ESRI ArcMap project, borehole Web Feature Services to post borehole logs in a 3-D mapping application, or water chemistry data Web Feature Service to bring data into a spreadsheet or database.

User searches for boreholes in an area. Returned metadata records have links to metadata for related resources, like logs of different types, core, water quality data, etc. that the user can follow to browse metadata for these resources.

A catalog operator wishes to import and cache catalog records from a collaborating catalog that have been inserted or updated during the last month (harvest). This operation requires knowledge of the metadata standard and version used for the returned records.

A user discovers an error in a metadata record for a resource that they have authored, and wishes to contact the metadata producer to request correction.

A search returns several results that appear to contain the desired content, and user must select the most likely to meet their needs. Metadata should provide sufficient information to guide this decision.

A project geologist at Company X is searching for data relevant to a new exploration target, and wishes to restrict the search to resources that are publicly available.

Complex search examples (see further discussion in the Query complexity section, below):

Search based on related resources, for example a search for boreholes that have core.

Boreholes that penetrate the Escabrosa formation.

Sample locations for samples with uranium-lead geochronologic data.

Find links to pdfs of publications by Harold Drewes on southeast Arizona.

Find geologic maps at scale < 100,000 in the Iron Mountains.

Who has a physical copy of USGS I-427?

Efficient searching

A search should return results that are actually relevant. Existing web search tools are very good at indexing relevance based on association of words in text, and using links and user navigation history for those links. This kind of indexing does not work for datasets, in which the information may be encoded in binary format, and proximity of strings may be a function of the data serialization algorithm, not the semantics. Semantic technology is advancing rapidly, and there is significant effort devoted to increasing search efficiency using background information (common sense) encoded in ontologies. To index structured data more effectively and take advantage of semantic technology, users must describe resources using controlled vocabularies (ideally linked to an ontology) in a formal metadata schema. Practically speaking, semantic technology is still in its infancy (maybe early childhood?), but the issue is important for discovery of structure data. Thus, use of controlled vocabularies for metadata content that is meant to enable search for particular resource characteristics is a requirement. Determining the elements requiring such vocabularies must be based on specific use cases.

Identifiers

A widely used identifier scheme is important to reduce duplication, and determine associations between resources. Globally unique identifiers are essential for the described resource, and for the metadata record.

The current thinking in the WWW community appears to be converging on a consensus to use HTTP URIs that are expected to dereference to some useful resource representation. A widely used and understood identifier scheme also enables semantic web functionality; "anyone can say anything about anything" requires being able to identify the things. Of primary interest here are crowd-sourced tagging of resources and feed back on utility, and related resources.

Query complexity

The complex search examples in the use cases section involve associations between resources, or resource-specific properties. The following table is a decomposition of some complex query examples

Table 1. Analysis of complex queries

Case#	Plain language query	Decomposition	Simplified solution
1	Boreholes that have core in a particular depth interval in a given area.	Borehole-centric approach-- geographic search for borehole resources (assume collar location), filter for those that have a related resource 'core', filter again for property of related resource 'core interval = min, max depth meters'. Alternatively, view search as actually for a 'core' resource, so search should be for 'core' with some given vertical extent. The core resource must provide an ID 'xxxx' for the borehole from which it was obtained. To obtain more details about the borehole, search for metadata on borehole with resource ID = 'xxxx'.	Include keywords for other resources associated with borehole. Put information about these in the abstract. User searches catalog for borehole with keyword (thesaurus=related resource) = 'core', reads abstract to see if its what they want. The keywords would have to be a controlled vocabulary.
2	Boreholes that penetrate the Escabrosa formation in a given area.	Geographic search for borehole resources (assume collar location), filter for property 'intersects Escabrosa formation'. Alternatively, search for borehole service that includes property = "formation tops", then query that service. Service properties would have to be from controlled vocabulary.	Include names of penetrated formations as keywords on a borehole. Formation names ideally from a geologic unit lexicon.
3	Locations for samples with uranium-lead geochronologic data in a given area.	Search catalog for Geochronology data service with property = 'analysis type' and backtrack to location point through sample metadata, or search catalog for U-Pb Geochronology Data Service and backtrack to location point through sample metadata, or search for 'sample service' with property = 'analysis type'. In the second case, there would still need to be some metadata property to indicate the analysis type for the service. Ap-	Include keywords for kinds of analytical data associated with a sample in the sample metadata record. Search for samples with keyword (thesaurus=analysis type) = 'U-Pb geochronology'.

		proach via the analytical data service requires chaining to the sample feature service, analogous to case 1 for borehole service.	
4	Find links to pdfs of publications by Harold Drewes on southeast Arizona.	Search for document resource with author = 'Harold Drewes' and geographic extent = 'SE Arizona', and online distribution format = 'pdf'.	Is search by representation format high enough priority to support?
5	Find geologic maps at scale < 100,000 in the Iron Mountains.	Search for geologic map resource with geographic extent = 'Iron Mountains, and resolution scale denominator < 100000.	Is search by resolution high enough priority to support
6	Who has a physical copy of USGS I-427?	Search for document publisher = USGS, Series ID = I-427, offline distribution format = 'paper copy'	

Consideration of these queries indicates a requirement to distinguish metadata service from a data service. When the request involves properties of specific instances of a particular resource type, a data service for that resource should be accessed. The metadata for that service should describe the properties offered for resource instances in that service.

Cases 1-3 can be handled in a general way by a service chaining process, in which the catalog is searched for services offering the feature of interest with the property of interest that will be used as a selection criteria. This approach keeps the top level resource catalog simpler, but makes discovery operations significantly more complex. Cases 1-3 can also be handled with scoped keyword terms, where the scope includes things like 'analysis type', 'geologic unit', 'related resource type'. In this usage, the scope specifies a controlled vocabulary of categories related to some concept. Addition of new querying capabilities requires adding additional scoped keywords in the metadata. The second approach is viewed as more appropriate in a 'keep it simple' design framework for minimum metadata requirements.

Cases 4-6 are related to document-oriented searches, for which distribution format and online access are important, and a number of bibliographic properties (scale, publisher, series, series ID, media, file format) come into play.

Accessing resources

Strong conventions for what kind of URL's are in metadata and how they are typed so that software can utilize them without operator intervention. Links in metadata to access resource should in general be complete URL's that can be invoked with a simple HTTP get, without having to add additional request parameters. Formal elements (with controlled vocabulary content) should provide machine-processable information to distinguish links that will return a document from links that invoke a service or access an online interactive application. The idea is that sufficient information should be provided that client software can parse the metadata record and provide useful functionality on the resource with minimal user interaction.

For many resources, different representations may be available. These might be different file formats for the same document for information resources. For non-information resources, a variety of representations that have different uses might be available. For example a physical sample may be represented by a text description of the sample, a GeoSciML xml description, visible light photograph, or images of the sample using other sensors. A geologic map may be available as a paper copy, a scanned image, a georeferenced scanned image, a vector data set in one of several formats (gml, shape file, file geodatabase, MIF, DWG), through a web map service, or through a web feature service. Metadata for a resource should be able to describe all of these different representations that the resource provider wishes to make available, in such a way that automated clients can seek representations useful to that client, or search clients can present users with links to access different formats or representations.

Citation and contact information

Citation information specifies the source of some content. Citations for the described resource specify the source for the resource intellectual content. The cited agent may have played various roles relative to the resource—author, compiler, editor, collector etc., and a controlled vocabulary is necessary to specify these. Citation for a metadata record specifies the agent responsible for producing the record, typically thought of as the metadata record creator. Metadata production involves elements of authoring, compiling, and editing. Minimally, citations must identify an individual person, an organization, or a role in an organization that is the agent filling a specified role relative to the cited resource. In most cases an organization will be specified, either as the employer or sponsor of a person, an institutional actor, or the host for some role (web master, metadata editor). In addition, information required to contact the cited actor is required to enable metadata users to contact a person with some knowledge of the cited resource. For long-lived metadata, contact for an agency role is most likely to persist. The minimum contact information required is either an e-mail address or telephone number.

Fitness for purpose

The metadata should provide sufficient description of the resource for a user to determine if the resource is likely to meet their needs, and to determine what representation to access. At the simplest level, such information should be provided in the abstract in the metadata record. This puts the onus on metadata producers to document in the abstract information that will be useful for users to determine fitness. Such information includes why the resource was produced, what sort of observation procedures were used, assessment of data completeness, accuracy, and precision, and comparison with other known similar resources. The data quality section of ISO19115 provides a data structure to formally describe this information, but the cost of using this is high (complex data entry), and there do not currently appear to be clients that utilize the information. The guiding principal should be that if users need to search on some particular quality criteria, specific guidance on how to encode that criteria in the metadata (which ISO19139 elements, what controlled vocabulary to use if terminology is involved) is necessary. This is out of scope for a minimum metadata requirement.

Branding

In a distributed, federated catalog system with harvesting, metadata records are expected to propagate far beyond their original point of introduction into the system. If an organization producing metadata wishes to be recognized, and in order for users to be able to contact the metadata originator, contact information for the metadata originator must be considered part of the metadata record, and maintained in harvest processes. For presentation to users, it is desirable to provide a link to an icon that can be displayed with records to brand the origination of the metadata.

The same considerations hold for the resource itself.

Access constraints, legal limitations

Metadata records that are not for public consumption should never be exposed to a harvesting request. Implementation of security and access control must occur at a lower layer in the network stack than the catalog service is operating, such that authorization/ user authentication information is handled by the environment containing the catalog client and server. Metadata for commercially licensed resources may be publicly accessible, but should clearly indicate the licensing requirements and procedure.

Low cost of entry

Metadata producers should be able to reuse and build on existing structured metadata. Minimum requirements should be limited to information that is commonly available. Resource specific details should be provided in text elements in the metadata. Special information necessary to utilize web links (e.g. web service operation) in metadata should be provided by text in the metadata or through linked documents.

Content requirement

Based on the above discussion, the following content requirements are specified.

1. Title of resource
2. Text description of resource. Free text, any length.
3. Globally unique identifier for resource
4. Citation for creator of resource (may be person, organization, or role).
5. Contact information for creator of resource (consider e-mail address to be minimum required)
6. Keyword for kind of resource described (from controlled vocabulary). At least one, could be many.
7. Bounding box, if applicable. For interoperability, proscribe geographic coordinates, WGS 84
8. Information necessary to access resource. This depends on whether distribution is physical or electronic. In either case, a text description of how the access works should be provided. At least one access process required.
    a. Physical distribution — applies to resources that are accessed by acquiring (or viewing) some physical artifact. This may be a rock sample, paper document, or physical media containing electronic information like a floppy disk. Requires text instructions on how to access, contact name, contact address. Other representations may provide limited ac-

cess via electronic communication (e.g. an image of the sample), these would be described using process for electronic distribution.

    b. Electronic distribution – access method, URL with term from controlled vocabulary specifying what function an HTTP get using the URL will invoke, and a controlled vocabulary term specifying the format (MIME media type) of a file-based response if applicable. The function might be return an html page, and electronic document in some other format, an end point for a service, an online application that requires user interaction, etc.

9. Globally unique identifier for metadata record
10. Date of last update of metadata record (create date if no updates have been made)
11. Metadata standard name, standard version, profile used, profile version (include all that apply, using controlled vocabulary).
12. Citation for creator of metadata
13. Contact information for creator of metadata

Information that will be assumed unless specified otherwise

1. character encoding in metadata
2. language of metadata (English)
3. language of resource (English)

Resource specific requirements

1. for resources that pertain to the subsurface, ocean, or atmosphere, a vertical extent. Recommend extent in meters, measured positive up from mean sea level. Depth extent requires specification of surface elevation, and depth (positive down) in meters. These two conventions would have to be distinguished using a coordinate reference system identifier.
2. for published documents standard bibliographic information—publisher, series, volume, page numbers
3. for spatial data specification of spatial resolution, terms to categorize spatial representation type (raster, polygon, lines)
4. for services,
    a. service type from controlled vocabulary,
    b. URL for service-specific document that describes operation of service (e.g. OGC GetCapabilities, WSDL)
    c. base URL for service requests
    d. contact information for service provider

Optional but highly recommended

1. Citations for resource creator and metadata creator should include URL for icons to display to brand content in presentation to user.
2. Scoped keywords from community thesauri to increase search efficiency. A gazetteer thesaurus like USGS place names is one obvious candidate. Details need to be determined.

Encoding

ISO19139

FGDC

Dublin Core CSW record

References

Cited literature

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

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

[ANZLIC, 2007] ANZLIC Metadata Profile Guidelines, Version 1.0: Turner, ACT, ANZLIC - the Spatial Information Council, ISBN: 978-0-646-46940-9, 372 p.

[INSPIRE ISO19115/119] Drafting Team Metadata and European Commision Joint Research Centre, 2009-02-18, INSPIRE Metadata Implementing Rules: Technical Guidelines based on EN ISO 19115 and EN ISO 19119,v. 1.1: European Commission Joint Research Centre, MD_IR_and_ISO_20090218.

Appendix K. Data items of interest for National Geothermal Data System [4-17-2010]

Table 2. Data items of interest for National Geothermal Data System. Compiled by S. M. Richard from proposals by State Geological Surveys and phone conversations with NGDS Boise State project partners.

Data Item	Item count	Services	Notes
Metadata	1170000	CSW, ISO19139	Will have to integrate metadata generated using a variety of specifications.
Borehole temperature data	570000	WFS: observation; WCS for continuous; WITSML log	Different kinds of temperature observations may need some clarification. Use coverage service for continuous temperature logs.
Borehole lithology log	94800	WFS: mapped feature; WITSML log?	Treat as borehole features with described intervals modeled as GeoSciML mapped features. Investigate WITSML log data structure as alternative
Water source characterization	84800	CUAHSI observation; WFS:observation, chemical property	Water sample is sampling feature, but there are a wide variety of possible characteristics of interest, including chemical, physical, and flow-related. Thus a soft-typed approach with {property, Measured value} pairs is recommended. Use Water Chemistry Data Item as part of this service
Document	15500	Document repository	Catalog service used to discover documents using metadata that includes URL to retrieve document from repository. Repository management discipline will have to evolve as repository grows.
Developed geothermal system feature	13700	for documents, catalog service using location; ISO19139 metadata: WFS geothermal system feature	**Need to model a** 'geothermal system feature', determine what important characteristics to describe are.
Water Chemistry	7260	CUAHSI observation; WFS:observation, chemical property; EarthChem chemistry service	Probably would be included as facet of water source characterization, as well as separate services to report water chemistry and quality. May be able to use EarthChem service here as well.
Bottom hole temperature	5950	WFS: observation, temperature	
Heat flow measurement	2700	WFS: observation, heat flow	
Digital well log	1400	WFS: WITSML log; OGC coverage	Properties measured by log may also be observed property for observation instances associated with other kinds of sampling features (site, sample…)

Data Item	Item count	Services	Notes
Rock chemistry	1000	WFS:observation, chemical property; EarthChem chemistry service	Similar logical schema to water chemistry service, different constituents in analyses.
Hot spring description	650	for documents, catalog service using location; ISO19139 metadata; WFS: hot spring xml?	**Need to model** a 'Hot spring feature', determine what important characteristics to describe are. Note a hot spring may be considered a kind of Water source, thus this model and water source characterization model may be closely related.
Drill stem test	400	WFS: observation. WITSML	Need to investigate WITSML to learn if it has reusable parts for DST's
Text description	200	Document repository or content management system	These may be treated as documents or nodes in a content management system like Drupal, or simply package in a database? Need more information on precisely what the information items are.
Geologic Unit geothermal characterization	150	WFS: sample, GeologicUnit: physical properties	**Need to define** physical properties necessary for geothermal characterization that are not in GeoSciML.
Volcanic vent description	140	WFS: geologicFeature: vent	**Need to model** a 'Volcanic vent feature', determine what important characteristics to describe are.
Thermal conductivity measurement	100	WFS: observation: thermal conductivity	
Active Fault	100	WFS: GeoSciML fault	fault feature, what is required content
Flow rate	100	WFS: observation: flow rate	
Geologic map	50	WMS for scans, WFS for vector data; Document repository	File discovery from catalog and download from repository for file-based (shapefiles, MIF, e00, fileGeodatabase, GML) map representations.
Permeability	20	WFS:observation: permeability	
Enhanced geothermal system feature	20	WFS: Geothermal feature	**Need to model** a 'Enhanced geothermal system feature', determine what important characteristics to describe are.
Resource suitability map	10	WMS for scans, WFS for vector data; Document repository	**Need to model** a 'resource suitability feature'. File discovery from catalog and download from repository for file-based (shapefiles, MIF, e00, fileGeodatabase, GML)
Alteration description	5	WFS: sample, GeologicUnit:	GeoSciML provides properties for alteration. Feature is sample or geologic Unit.
Fluid inclusion data	5	WFS: observation: equilibration temp, etc	FLINC observation results data type

Data Item	Item count	Services	Notes
Geothermal map	2	WMS for scans, WFS for vector data; Document repository	Need to explore what defines such a map. File discovery from catalog and download from repository for file-based (shapefiles, MIF, e00, fileGeodatabase, GML)
Aquifer temperature map	1	WMS for scans, WCS--coverage, temperature/depth grid	
Crustal Stress data		WFS: observation, stress state tensor	stress measurement
Gravity data		WFS: observation, gravity station; WCS gravity grid; WMS for scanned map	
Intrusive body with heat		WFS: Geologic unit	Geologic unit is intrusive body, geologic unit description
Earthquake epicenter		WFS: mapped feature: 3D point, EQ epicenter feature	**Need to model** a 'earthquake epicenter feature'
Samples		WFS: sample; use SWE xml	In general, samples will need to be associated with other observation data.
Production statistics record		Energistics ProdML	**Need to model** a 'geothermal production record'. Energistics ProdML is concerned with production for oil and gas, may be able to extend easily for geothermal. Check CUAHSI for stuff on water well production. Guess that

Notes.

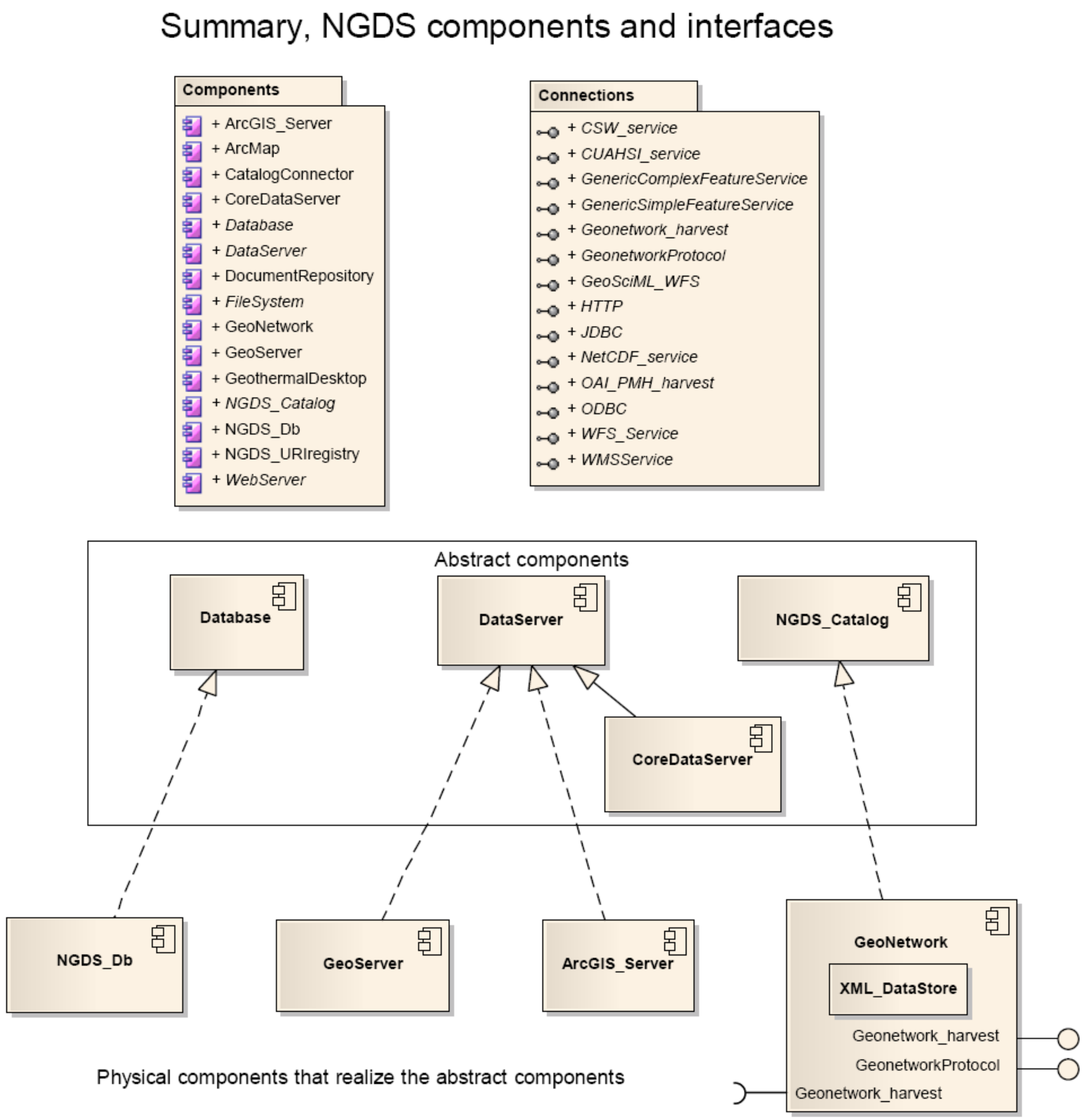
Data item is an information resource identified by a stake holder as an item of interest.

Item count is the approximate number of instances for that item projected in the system.

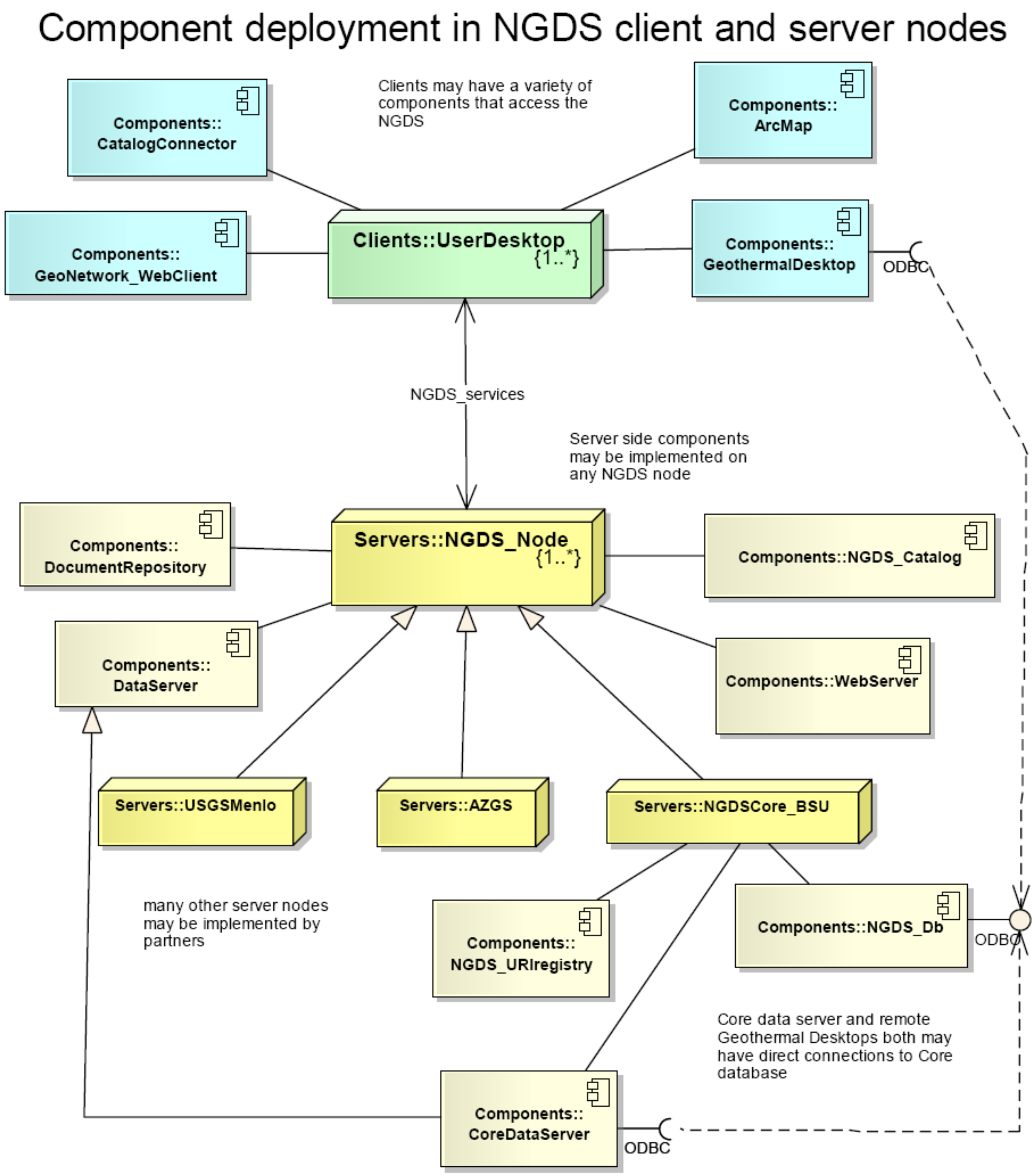
Services summarizes the existing service profiles and markup languages that may be used for a data item.

Appendix L. Summary, NGDS components and interfaces [4-20-2010]

UML sketches to promote discussion on system architecture and serve as starting point for system design documentation. Generated using Enterprise Architect v7.5, exported as TIFF images for placement in text document. An HTML version was also distributed that includes links from UML packages and classes to some explanatory material.



Appendix M. Component deployment in NGDS client and server nodes NGDS [4-20-2010]



Appendix N. NGDS-DOE Minimum Metadata Reporting Requirements [4-20-2010]

Distributed with request for statements of work from state partners in AASG Geothermal data project. Also proposed for consideration by NGDS Technical Working Group. This document superseded by Metadata Recommendations for Geoscience Information Resources, which is a revised version of the document presented in Appendix J of this document.

NGDS-DOE Minimum Metadata Reporting Requirements

Wolfgang Grunberg, AZGS

Metadata - data about data - is used to describe, discover and access digital or physical resources. The NGDS-DOE minimum metadata reporting requirements are designed to balance the need for on-line, interoperable metadata discovery and distribution with the cost of generating digital metadata. In order to effectively advertise your resources, metadata documents must accomplish three major goals:

1. **Describe** the digital or physical resource or service.
2. **Credit** the owner, author, or responsible party of the resource.
3. **Provide access** information to the described resource

Metadata Fields

Key: *Groupings;* **<u>required</u>** and **optional** metadata fields; (number of values that can be specified).

- *Citation Information*
    - **<u>Title</u>** (1 entry): Human readable.
    - **<u>Description</u>** (1 entry): Inform the reader about the resource's content as well as its context.
    - **<u>Author</u>** (1 to many entries): Authors, editors, or corporate authors.
    - **<u>Publication Date</u>** (1 entry): Use a "year", "day/month/year", or ISO 8601 date and time format.
    - **Keywords** (0 to many entries): Thematic, spatial and temporal free-form subject descriptors.
    - **Resource language** (0 to 1 entries): Use three letter ISO 639-2 language code. Defaults to "eng" for English.
    - **Resource ID** (0 to many entries): A resource identifier following any public or institutional standard.
    - *Citation Contact* (0 to 1 entry): The party responsible for creating the resource. **Organization name**, **person name, street address**, **city**, **state**, **ZIP code**, **email**, **phone**, **fax**.
    - **Bibliographic Citation** (0 to 1 entries): Enter a full bibliographic citation if the resource has been published.
- *Geographic Extent* (1 entry, minimum bounding rectangle or point): **north bounding latitude, <u>south bounding or point latitude</u>, east bounding longitude, <u>west bounding or point longitude.</u>** Values given in decimal degrees using a WGS84 datum.

- *Temporal Extent* -- enter geologic time periods as keywords
    - **Start date** (0 to 1 entries): Use ISO 8601 date and time format.
    - **End date** (0 to 1 entries; requires a start date): Use ISO 8601 date and time format.
- *Distribution Information*
    - **Link to the resource** (0 to 1 entries): A URL pointing to a resource or resource webpage.
    - **Access instructions** (0 to 1 entries): A sentence or paragraph describing how to access the information.
    - *Distribution Contact* (1 entry): The party to contact about accessing the resource. **Organ-ization name**, **person name, street address**, **city**, **state**, **ZIP code**, **email**, **phone**, **fax**.
- *Metadata Information*
    - **Metadata UUID** (0 to 1 entry): A Universally Unique Identifier (UUID) will be assigned to your metadata if it is not provided.
    - *Metadata Contact* (1 entry): The party to contact with questions about the metadata it-self. **Organization name**, **person name, street address**, **city**, **state**, **ZIP code**, **email**, **phone**, **fax**.

Appendix O. AASG-NGDS Statement of Work outline [4-20-2010]

AASG-NGDS Statement of Work outline

Introduction

The goal of this project is to get information useful for geothermal exploration, resource evaluation, and development, easily accessible online. Information from different data providers must be presented in a consistent, documented format. System users must be able to locate information anywhere in the system through a single search. Information resources must be accessible online. To meet these requirements, data produced for inclusion in the system must be accompanied by metadata describing datasets and how to access them. To enable interoperability, data sets will need to be published using an interchange format adopted by the community and documented for system use.

This effort hinges on development of a community of practice using a shared collection of protocols for publishing, finding, and delivering digital information online. One of the foundations for this community is a shared vocabulary for discussing the architecture of the system, the protocols used for communication, and the nature of the information resources themselves. Some of this vocabulary is introduced in this document. Like any community vocabulary, it is dynamic and expected to evolve as our community grows. Definitions included in the glossary are intended to help readers understand the terms as they are used here.

The Statement of Work

The AZGS has the task of assuring that data compiled by the subcontractors are integrated into the National Geothermal Data System. The goal is to expose data in one or more interchange formats for which software applications (particularly the geothermal desktop application in development at Boise State University under a parallel DOE project) are available that can read and utilize date in the interchange format directly. Compatibility with the interchange format will not dictate the logical and physical data models used for internal data management by the producing agency. What is necessary is that the information content in the internal data model can be mapped in a systematic way to the interchange format. To enable integration, partners need to confer with the AZGS team to determine the most effective mechanism for data integration.

For contractual, evaluation, and reporting purposes, specific metrics are required in the statement of work. In the accompanying spreadsheet, this is represented by the 'No. items to be entered' column for resources that are to be scanned or digitized, and for data that are already in a digital form, quantification by number of records or bytes in the 'Amount of digital data exposed to the NGDS' column.

The statement of work worksheet also requests information on the data delivery plan, and a statement of the significance of the proposed data to the NGDS. These factors will be considered by the SAB in evaluating the scope of work.

Data to be scanned

Reports, logs, maps and other documents pertinent to geothermal energy exploration, evaluation, and development that exist in hard copy but are not available online may be converted to digital form by scanning to create digital image files. If the resource is a map, it should be georeferenced (geoTiff or world file) if possible. Preferred document formats are pdf, tif, jpg, or png. File formats that are specific to particular software are undesirable. OCR processing of text to make Adobe Acrobat files searchable is highly desirable. Georeferenced map images ideally will be published through a web map service as well as accessed from document repositories.

Digital data

Data to be digitized

Geospatial and measured data are most useful in a computer analyzable format. Paper manuscript maps should be scanned and georeferenced to start. Maps that represent discrete map units or linear features (like a geologic map) can be converted to a vector GIS digital format if they are deemed to have sufficient value. Tabular data associated with sites or samples (e.g. chemical analyses, gravity data) may need to be manually converted from text to tabular digital data.

Existing digital data

Data that are in a structured digital format can be published for viewing using web map service, delivered as file-based GIS datasets, or published as a web feature service. The choice will depend on the nature of the dataset and the capabilities and priorities of the data provider. In order to make information interoperable, the system will develop conventions for data interchange formats that may require conversion of internal data formats and vocabularies to an interchange format. This will be the time-consuming part of publishing existing datasets. For resources that are digitized under the auspices of this project, crafting of the digital conversion process can make generation of the interchange formats easier. In some cases, the interchange formats will need to be developed based on the information that project partners wish to publish to the system.

Service availability

Data delivered by publishing as an online service must be guaranteed to be available 24/7 at least for the duration of the project (3 years). Part of the NGDS project is development and implementation of a business model for long-term maintenance of system resources.

Technical discussion

Data delivery options

- Register files in a document repository, submit metadata to catalog.
- Implement web service, either at your agency, a NGDS data center (Boise State, state geologic surveys--Kentucky, Illinois, or Nevada), or by arrangement with another agency. Submit metadata to catalog.

Data will be considered part of the NGDS when it is locatable using the NGDS core catalog, and accessible via the web according to procedures described in the metadata record obtained from the NGDS core catalog. The anticipated delivery process must be defined in the Scope of Work statement (Data delivery plan column in spreadsheet).

Metadata

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

Individual documents require one metadata record per document. The metadata must include the URL at which the document can be accessed. These documents might be scans of well logs, scanned reports or publications, or data in a spreadsheet, such as an Excel file.

Datasets include internal record level source information, documenting details of observation procedure and other information specific to a particular data type. This metadata is delivered with the data, and only summarized in the dataset metadata that is published to the NGDS catalog.

The required metadata content is explained in the accompanying 'Minimum metadata content requirements' document. We are not proscribing a particular data format for the metadata, but strongly recommend FGDC xml or ISO19139 xml. Please confer with the AZGS developer team about metadata formatting to facilitate import of metadata into the NGDS catalog.

Datasets:

The following Table 1 is a summary of the data items proposed for delivery via the National Geothermal Data System by project partners in the survey spread sheets submitted during proposal preparation. This list is dynamic; as the project evolves new data items will be added as necessary.

The term 'Data item' used here to denote a kind of information that has a well-defined scope, which may be defined by the physical property specified, or by the kind of information artifact involved. The purpose of this classification is to analyze the information of interest into types that will use the same data schema, delivery mechanisms, and metadata schema.

A number of the data items in Table 1 will require additional modeling to determine a useful collection of attributes to specify delivery of interoperable data. Our plan is to confer with partners proposing to deliver those data items to develop a model that includes the necessary and available information. In order to make data specifying physical properties interoperable, we will be developing requirements for content of data records, and the units of measurement used for reporting.

The category column in Table 1 groups data items into higher-level information categories that correspond to the 4 main categories of services in the system architecture. These are:

- Observation – an information resource representing the event of observing and recording properties of some feature (Open Geospatial Consortium, Observations and Measurements, http://www.opengeospatial.org/standards/om). Includes: a result, which is the measured or observed value; a feature of interest, which is the feature the observer wishes to characterize; a procedure, which includes information on who made the observation and how it was made. A sampling feature may be specified to record what part of the feature of interest was the actual target of the observation. Observations represent the basic data that are the foundation for the other information categories. The observation model allows modeling composite observations, which may represent the aggregation and interpretation of one or more input observations.

- Feature – an information resource representing some identifiable thing of interest in the world. A feature is described by a collection of attributes that are typically each the result of one or more observations. Features present a more aggregated or interpretive view of the world than observations (although a feature can be modeled as the result of an observation). Features will be delivered via OGC Web Feature services (Open Geospatial Consortium, Web Feature Service, http://www.opengeospatial.org/standards/wfs) or other similar services. Typically, features have a geographic location.
- Document -- A packaged body of intellectual work; has an author, title, some status with respect to review/authority/quality. USGS peer reviewed would be a 'status property'. Have to account for gray literature, unpublished documents, etc. A document may have a variety of physical manifestations (pdf file, hard-bound book, tiff scan, Word processor document...), and versions may exist as the document is traced through some publication process. May be map, vector graphics, text. Sound, moving images are included as document types.
- Coverage – A dataset that reports the values of some continuously varying property over some spatial extent. Examples include well logs that report the values of resistivity, density, or some other property along the well bore, gravity maps that report measured (perhaps by extrapolation) values of gravity over some geographic region. A coverage may be the result of one or more observations. A coverage may also be associated with an individual feature, such as a map showing the thickness distribution of a geologic unit or the average temperature at some depth in a geothermal system.

The discussion of data items in table 1 assumes familiarity with the basic Open Geosptial consortium (OGC) service architecture. GeologicUnit and GeologicStructure are used as defined for the GeoSciML xml markup language, and system implementers are strongly encouraged to become familiar with that specification as well.

Table 3. Data items for NGDS.

Data Item	Category	Notes for data product
Active Fault	Feature	A GeoSciML GeologicStructure feature; attributes should include at least a statement of evidence for fault being active; ideally includes orientation information, time since last displacement, hydrologic information about fault zone if available.
Aquifer temperature map	Document	A map document.

Data Item	Category	Notes for data product
Borehole lithology log dataset	Coverage	Lithology log consists of collections of intervals defined by top and bottom coordinate in borehole trace, and association with lithology description. Lithology descriptions will include original recorded text, lithology categories from CGI vocabulary, and other properties. We anticipate using the GeoSciML schema (http://geosciml .org) for these descriptions, with extensions if necessary. Log will be associated with a borehole collar location, and metadata for the original description. A paper copy of a log or a scan of a paper log is considered a kind of document.
Bottom hole temperature	Observation	See Temperature data
Chemical analysis (whole-rock chemistry)	Observation	Individual records will be a collection of {measured constituent, abundance pairs}, with identifiers for the analyzed sample, and analysis procedure. The procedure is considered to include the who, how, with what equipment information for the analysis.
Crustal Stress data	Observation	Associated with a site. Needs model
Digital well log	Coverage	A dataset that consists of a collection of measurements of some physical property as a function of depth in a borehole. WITSML (Energistics), NetCDF, LAS are possible interchange formats.
Document	Document	Document is used to mean a packaged unit of content with a single authorship (which may include several people). Examples include books, reports, journal articles, geologic maps, other kinds of maps. Internal content within a document is generally not individually identifiable (unlike records in a dataset).
Drill stem test	Observation	An observation feature that includes the results of a drill stem test. Needs modeling of key observation results, including pressure, fluid composition.
Earthquake epicenter	Observation	Treat as observation because epicenter location is always the result of a measurement and analysis process; is observation with result that depends on a collection of seismometer recordings. The epicenter can also be conceived as a feature, with the observation as metadata for definition of the feature.
Flow rate	Observation	Always an attribute of a water source feature or of a water channel feature.
Fluid inclusion data	Observation	Associated with a sample; needs content model
Geologic map	Document	Geologic maps will be made available through one or more of several mechanism: Download of image file (tiff, Jpg, or pdf), ideally georeferenced; an OGC Web map service based on a map image (no getFeatureInfo) or better on vector data; or as a vector data set. NGDS AASG project personnel can be assigned to assist with implementing map services, but requests for such assistance should be made in the work plan so we can schedule resources. Map services may be hosted by NGDS regional hubs, NGDS core, by the data provider, or any other reliable server, and will be required to be maintained available. Map images as files may be published in repositories

Data Item	Category	Notes for data product
Geologic Unit feature, geothermal characterization	Feature	A geologic unit description specifying properties important for geothermal energy evaluation; includes standard aquifer properties like lithology, permeability, porosity, as well as thermal properties like thermal conductivity and specific heat.
Geologic unit feature, Alteration description	Feature	Is GeoSciML alteration description sufficient, or should it be extended
Geothermal map	Document	see geologic map
Geothermal system feature	Feature	Data modeling is still necessary to determine a collection of attributes to characterize a geothermal system as a feature. Subtypes might include developed geothermal systems,
Geothermal system feature, Enhanced	Feature	Needs to be modeled, another subtype of geothermal system?
Gravity map data	Document	See geologic map. Spatial data may be grid or contours.
Gravity Station data	Observation	Associated with a site. See PACES for model.
Heat flow measurement	Observation	Heat flow measurements are based on a temperature gradient measured over some interval, and a thermal conductivity value for the material between the two temperature points. The location and temperatures defining the gradient and estimated conductivity must be reported in a complete heat flow report, along with procedure metadata.
Hot spring description	Feature	A hot spring is a kind of water source, which is required to have temperature data for the water produced, along with other properties associated with a water source (location, flow rate, water chemistry data, time series for flow and chemistry?). Modeling still necessary.
Intrusive body with heat	Feature	Treat as GeoSciML Geologic feature with geothermal characterization
Permeability	Observation	May be reported through observation service associating individual samples with permeability measurements. Permeability may also be reported associated with a geologic unit in a GeologicUnit geothermal characterization.
Production statistics record		Needs modeling. Tabular time series data may require a different service category. Has geographic location associated with producing system, so could be considered a Feature or Observation.
Resource suitability map	Document	Map document. Spatial data could be coverage or polygons.
Sample	Feature	Associated with site; becomes sampling frame for variety of other observations
Temperature data	Observation	Attributes will include temperature, units, X, Y, Z coordinate, borehole identifier, and measurement procedure. Different measurement procedures will need to be documented. Bottom hole temperature data is one kind of borehole temperature data. Temperature log datasets (as opposed to scanned log documents) are treated as a kind of digital well log. Typically is measured in a borehole to be geothermally interesting.

Data Item	Category	Notes for data product
Thermal conductivity measurement	Observation	Observation feature, attributes include identification of sample used for measurement, procedure, and result with uncertainty.
Trace constituent chemistry dataset. (Water Chemistry, trace-element data)	Observation	Trace element chemical analyses report concentrations of constituents that do not form a significant part of the total material. Most water quality or water chemistry data fall in this category, as well as rock trace-element data. Individual records will be a collection of {measured constituent, concentration pairs}, with identifiers for the analyzed sample, and analysis procedure. The procedure is considered to include the who, how, and with what equipment information for the analysis.
Volcanic vent feature	Feature	Volcanic vent has a location that may be represented by a point or polygon. Has additional properties that need to be modeled more completely. This will need to be done by partners wishing to contribute this sort of data. A complex data model could be imagined, including eruption history, magma composition, fluid and gas compositions, eruption rates, associated heat flow measurements, associated magma body… For NGDS purposes, we need to identify the key properties of interest as a starting point.
Water source feature	Feature	there are a wide variety of possible characteristics of interest, including chemical, physical, and flow-related properties. Thus a soft-typed approach with {property, measured value} pairs is recommended. We will confer with partners providing this kind of information to establish practices for common kinds of characterization. We anticipate being able to use CUASHI practices for much of this kind of data.

Glossary:

Aquifer: A geologic unit that is a hydrologically connected body of material.

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

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

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

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

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

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

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

Feature type: Type for representing a feature

Geologic structure: A configuration of matter in the Earth based on describable inhomogeneity, pattern, or fracture in an Earth material The identity of a GeologicStructure is independent of the material that is the substrate for the structure. GeologicStructures are more likely to be found in, and are more persistent in, consolidated materials than in unconsolidated materials. Properties like "clast-supported", "matrix-supported", and "graded bed" that do not involve orientation are considered kinds of GeologicStructure because they depend on the configuration of parts of a rock body. Includes: sedimentary structures. (from http://www.geosciml.org/geosciml/2.0/doc/GeoSciML/GeologicStructure/GeologicStructure.html)

Geologic unit: a body of material in the Earth whose complete and precise extent is inferred to exist (NADM GeologicUnit, Stratigraphic unit in sense of NACSN or International Stratigraphic Code), or a classifier used to characterize parts of the Earth (e.g. lithologic map unit like 'granitic rock' or 'alluvial deposit', surficial units like 'till' or 'old alluvium'). (from http://www.geosciml.org/geosciml/2.0/doc/GeoSciML/GeologicUnit/GeologicUnit.html)

Geothermal system: Need definition.

Information resource: A resource that can be transmitted electronically.

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

Representation: A binding between a symbol (in language, text, graphics, computer bits, etc.) and a human concept.

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

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

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

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

Recommended reading:

GeoSciML documentation, Available at http://www.geosciml.org/geosciml/2.0/doc/. Project home page is http://www.cgi-iugs.org/tech_collaboration/geosciml.html.

International Organization for Standardization (ISO), Geographic Information—Metadata: ISO 19115. Official site is http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020, but the document costs 224 Swiss Francs from there. Get it from http://webstore.ansi.org/RecordDetail.aspx?sku=INCITS%2fISO+19115-2003 for US$ 30.00.

Open Geospatial Consortium, Catalogue Service Implementation specification (CSW). Available at http://www.opengeospatial.org/standards/specifications/catalog. V2.0.2 is in use; See also the ISO Metadata Application profile, accessible from the same web page.

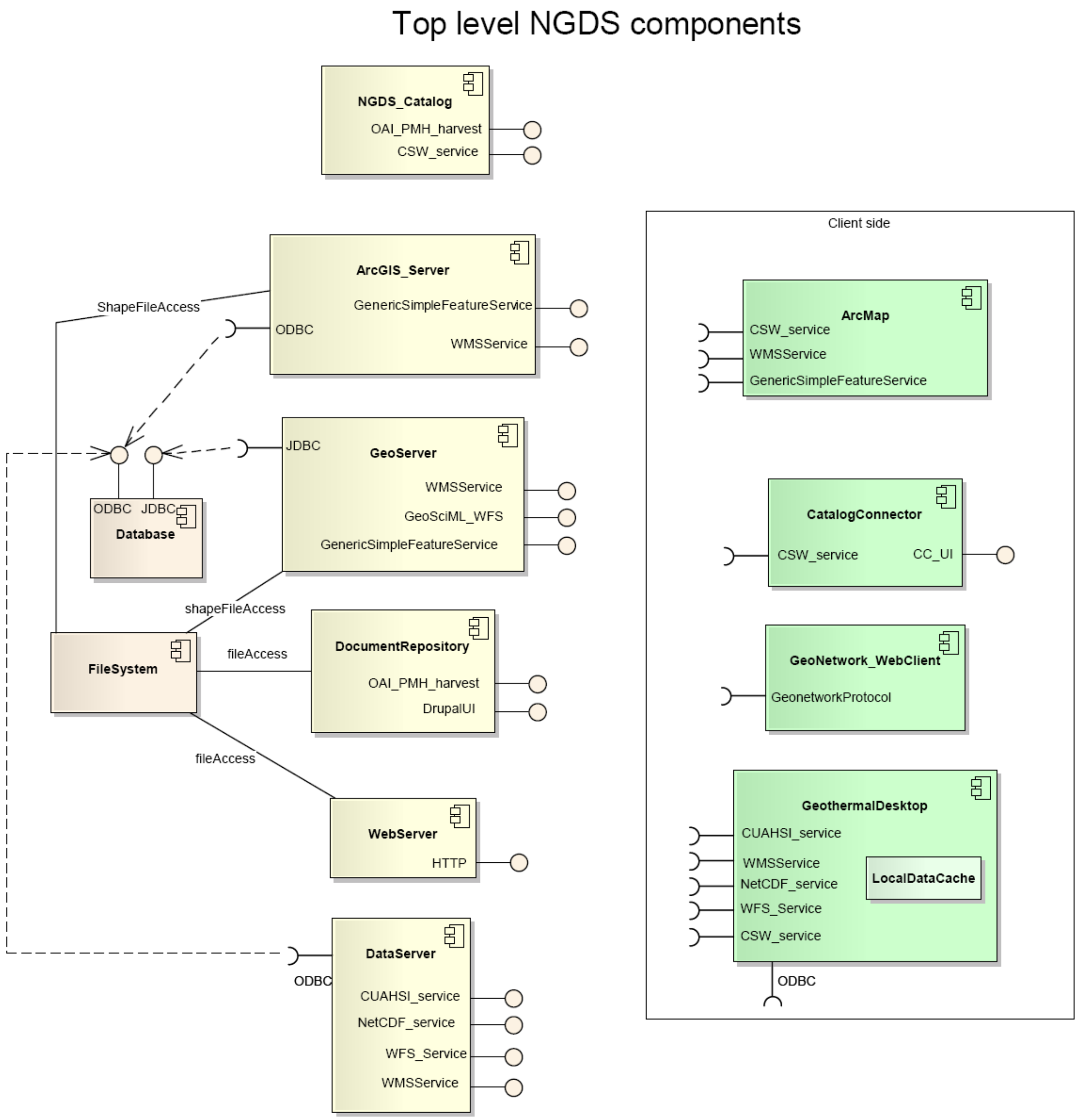
Open Geospatial Consortium, Observations and Measurements (O&M). Available at http://www.opengeospatial.org/standards/om. This has been adopted as ISO 19156.

Open Geospatial Consortium, Web Feature Service Implementation specification. Available at http://www.opengeospatial.org/standards/wfs. V. 1.0.0 and1.1.0 are in use; version 2 is expected soon with some significant improvements.

Open Geospatial Consortium, Web Map Server Implementation Specification (WMS). Available at http://www.opengeospatial.org/standards/wms. Versions 1.1.1 and 1.3.0 are in use.

USGIN ISO metadata guidelines, Available at http://lab.usgin.org/profiles/usgin-iso-19139-profile.

Appendix P. Top level NGDS components [4-20-2010]



Appendix Q. Heat flow data demonstration plan [4-22-2010]

Subject: Heat flow data demonstration

From: Stephen M Richard <steve.richard@azgs.az.gov>

Date: Thu, 22 Apr 2010 14:07:22 -0700

To: "Moore, Joe" <jmoore@egi.utah.edu>, Walter Snyder <wsnyder@boisestate.edu>, "Lund, John" <john.lund@oit.edu>, "Horne, Roland" <horne@stanford.edu>, Lee Allison <lee.allison@azgs.az.gov>, "Peterson, Jack" <jack_g_peterson@blm.gov>, "Williams, Colin" <colin@usgs.gov>, Lisa A Shevenell <lisaas@unr.edu>, Jordan T Hastings <hastings@unr.edu>, "Boyd, Toni" <toni.boyd@oit.edu>, "Jennejohn, Dan" <danj@geo-energy.org>, Kim Kurz <kimkurz@boisestate.edu>, Kewen Li <kewenli@stanford.edu>, "Anderson, Arlene" <Arlene.Anderson@ee.doe.gov>, Catherine Martinez-Wells <catherine.wells@azgs.az.gov>, "Gowda, Varun" <vgowda@egi.utah.edu>, "Cuyler, David" <dscuyle@sandia.gov>, "Bowers, John" <jsbower@sandia.gov>, "Jacobi, Melissa" <melissa.jacobi@go.doe.gov>, Gary L Johnson <glj@unr.edu>, "Logsdon, Grant" <grant.logsdon@go.doe.gov>, "Mink, Roy" <h2oguy@copper.net>, Rachael Willis <rwillis@egi.utah.edu>, "Reed, Marshall" <mjreed@usgs.gov>, Robin Penfield <rpenfield@unr.edu>, Charlla Adams <cadams2@boisestate.edu>, Daniella Morgos <daniellamorgos@boisestate.edu>

CC: "Bruce.Simons@dpi.vic.gov.au" <Bruce.Simons@dpi.vic.gov.au>

As we discussed in the phone conference yesterday, we'd like to spin up a demonstration web service for heat flow measurements. If any of you have heat flow data that's in a tabular form (spreadsheet, database..) and would like to participate, please e-mail me with information about what you've got--what form is the data in, what fields do you have in the heat flow records, how many records are there.

From the investigation I've done so far it looks properties that should be reported include:

Display Name (short text to label on maps)
Description (free text notes on the measurement)
Borehole identifier
date drilled
collar elevation
water table depth
gradient interval start depth
gradient interval end depth
corrected gradient
corrected heat flow
corrected heat flow error
thermal conductivity
source of data text field with better information on provenance

Please add to this list--we want to include information that 1. is useful, and 2. is available.

We have a quick and dirty service set up with the AZ data from the SMU geothermal data site. http://services.azgs.az.gov/ArcGIS/rest/services/AzGeothermal/MapServer describes the various service options from the ESRI point of view. I'm attaching an ArcGIS 9.3 layer file you should be able to load in ArcMap to see a view of the heat flow data.

This will be a demonstration of the system. The objective would be to have multiple data providers implementing the same service (OGC WFS, same xml schema for results) to provide all heat flow data that we as a community know about. This will provide a data source for testing Geothermal Desktop functionality, as well as other service clients. I suspect it will also highlight some of the challenges we face--'branding', providing provenance information to users, and detecting data duplication, among others.

Hope to hear from you soon

thanks
steve

Stephen M. Richard
Section Chief, Geoinformatics
Arizona Geological Survey
416 W. Congress St., #100
Tucson, Arizona, 85701 USA

Appendix R. Outline for an NGDS Technical Coordinating Council [5-26-2010]

DRAFT

Outline for an NGDS Technical Coordinating Council

Proposal

Establish a Technical Coordinating Council (TCC) among the projects that are funded to build and deploy the National Geothermal Data System

Purpose

To coordinate activities of NGDS development projects lead by Boise State University, the Arizona Geological Survey, Southern Methodist University, and the U.S Geological Survey in order to ensure efficient and effective design, deployment, and population of the NGDS.

Tasks

Promulgate plans and make recommendations to project managers on technical aspects of NGDS, including:

- Use cases to guide system development
- Priorities for technical development based on adopted use cases
- Adoption of specifications, including data and metadata content models, interchange formats, and service profiles
- Adoption of controlled vocabularies and ontologies for data integration
- Adoption of conformance and quality assurance tests and metrics

Identify possible duplication of effort or incompatible component development and recommend solutions to maximize development efficiency and effectiveness

Identify emerging best practices

Authority

The TCC is intended to promote agreement and consensus on technical issues facing NGDS.

Recommendations will be voted on by the voting members, and adopted by simple majority of the membership.

AZGS anticipates adopting the TCC as Technical Advisory Board for its project.

Recommendations of the TCC are not binding without agreement by affected project PIs.

Membership Voting membership includes one technically qualified representative from each project engaged in building or deploying NGDS. Proposed membership:

- Boise State University - Christian Loepp
- Arizona Geological Survey – Stephen Richard
- U.S. Geological Survey – tbd
- Southern Methodist University/Siemens – Fabian Moercher
- U.S. Dept. of Energy/Sandia National Laboratory – David Cuyler [Chair]

Additional non-voting members from the participating institutions may participate in TCC discussions, serve on subcommittees or working groups, or provide support to TCC. Non-voting members from outside projects may participate in committee or subcommittee discussions.

Working process The TCC will maintain a shared development calendar accessible to council members, project PIs, DOE program managers, and others at the discretion of project PI's. Each council member will keep the calendar up to date, showing current development activity, development target dates, and future development plans as they evolve. The TCC will schedule telephone conferences of the members as necessary. Voting on recommendations will be recorded and documented by posting on the geothermaldata.org web site.

Recommendations adopted by the council will be posted on www.geothermaldata.org. PI's intending to adopt recommendations will notify the community by adding a statement to that effect with the posted recommendation on www.geothermaldata.org.

Appendix S. Draft Phase 1 report: AZGS Contribution to System Design for National Geothermal Data System [8-21-2010]

Phase 1 report, Draft

AZGS Contribution to System Design for National Geothermal Data System

11/12/2010 9:08 AM

Introduction

This report is a contribution to design of the National Geothermal Data System (NGDS). The major subcontractor roles of Arizona Geological Survey (AzGS; representing the Geoscience Information Network) in this project, as stated in the Attachment to DOE Agreement #014G106215-A are "major participant in design of NGDS database (Tasks 1.0 and 4.0); lead role on design, standards and protocols associated with web services, development of data catalogs, and helping implement web services…" (Task 2.0). This report summarizes recommendations, many of which were presented and/or discussed with the project team throughout Phase 1 of the project, for system architecture particularly focused on the data publication and data access aspects of the system relevant to our role in the project. These recommendations are based on development work done at the Arizona Geological Survey under auspices of DOE award DE-EE0001120, in conjunction with related work on development of the Geoscience Information Network (US GIN) supported by NSF grant EAR-0753154, and a parallel DOE award DE-EE1002850 to compile and publish geothermal data from state geological surveys to integrate with the NGDS.

The recommendations in this report are by no means intended to be authoritative statements of the system architecture and data acquisition plan. They are based on considerable development work over the last several years under the auspices of a variety of other geoinformatics projects, and are intended to present an incremental development framework that utilizes existing technology wherever possible, builds on a variety of existing standards and specifications, and allows for agile development of the NGDS in the current, rapidly evolving technology environment. These recommendations are a starting point for system design and evolution.

This document includes an introductory section discussing the scope of the system based on the original FOA and the consortium's proposal and some use cases helpful to think about requirements. The second section outlines the architecture for distributed data access in the system. The third section discusses data acquisition, and a final section consists of some technical discussion and a summary of recommendations.

Scope and purpose of system

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

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

From http://apps1.eere.energy.gov/geothermal/projects/projects.cfm/ProjectID=27:
"The NGDS will be able to handle the full range of geoscience and engineering data pertinent to geothermal resources as well as incorporate data from the full suite of geothermal resource types. It will be able to handle data on geothermal site attributes, power plants, environmental factors, policy and procedure data, and institutional barriers. It will provide resource classification and financial risk assessment tools to help encourage the development of more geothermal resources by industry. It will be an easy to use system that meets the needs of the professional and the public for information on geothermal resources."

Abstracted from Original Project Proposal from the Geothermal Data Coalition:
Goal: build a state-of-the art data system.

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

System Technical Design principles

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

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

Users should be able to search all resources in the system through a single, but not sole, search client. Any search client that implements the system catalog service profile should be able to conduct search against any system catalog that also implements the profile. This means that there can be multiple portals and client applications for accessing system resources; it requires that a single client can search different catalogs in the system without the user having to reconfigure the software.

Providing quality information to evaluate system resources requires criteria that can be used to filter data and categorize them according to established and user-defined quality levels. These quality filters will vary depending on the type of data and their targeted use.

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

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

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

- Easily maintained

Data Access

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

Requirements

Use cases

Data access use cases

As a starting point for design of the NGDS, it is important to define the function of the system. The approach taken is to present a number of user scenarios or use cases that describe the kind of interaction envisioned for users of the system. This list includes a number of initial use cases collected in a brainstorming session at the kick off meeting in Boise, from the original project proposal.

- Get features that locate and describe exploration leases in a particular area defined in the user interface.
- Get a map image to add to the user map display that shows all boreholes drilled for a particular purpose (geothermal exploration, fluid injection, geothermal fluid production…).
- Get features for borehole collar locations selected based on kinds of information obtained from the boreholes (e.g. neutron density log, core, temperature measurement)
- Get a map image to add to a user map display showing borehole bottom-hole temperature and depth, plotted at the bottom-hole location (x,y,z).
- Get borehole interval feature with measured temperature gradient for that interval (z_1, z_2, gradient, collar location)
- Get geothermal spring features with location, a standard set of fluid chemistry data, flow data, and salinity
- Get all the data for an area of interest and make it accessible in a user workspace that can be saved for later use. User should be able to collect data from within a single application. Data integration from different sources should be transparent to user.
- Publish a data set to the system, creating metadata and making data set available for other users. System must provide documentation for procedures, and guidance on precision, units and formatting.
- Calculate financial risk based on weighted properties of geothermal features in a prospective area, along with any other significant factors.
- Adjust geothermal classification criteria, factoring in data quality (based on metadata) assigned to input for classification.

Metadata and catalog use cases

This section includes a number of user scenarios based on those compiled for the USGIN metadata recommendations for this project (USGIN Specifications Drafting Team and NGDS developer team, 2010/07/28). The fundamental use case addressed by a catalog system is to find resources of interest via the internet, based on criteria of topic, place, or time, evaluate resources for an intended purpose, and learn how to access those resources. Detailed metadata describing a resource data schema, describing service or application operation, or providing detailed descriptions of analytical techniques and parameter are outside the scope intended for basic search and discovery metadata. Our contention is that this more domain/resource specific type information is better accounted for with linked documents utilizing schema appropriate to those specific resources. Some examples include OGC getCapabilities, WSDL, and ISO 19110 feature catalogs:

- Find all documents related to a particular topic in any repository in the system.
- Find an online version of a map showing temperature gradient and include it as a layer in a project map visualization.
- A user specifies a geographic bounding box or one or more text keywords to constrain the resources of interest, and searches a metadata catalog using these criteria. The user is presented with a web page containing a list of resources that meet the criteria, with links for each resource that provide additional detailed metadata, and direct access to the resource if an online version is accessible, e.g. as a web page, Adobe Acrobat document, or online application (see Accessing Resources, below).
- A client application provides user with a map window that contains some simple base map information (political boundaries, major roads and rivers). User wishes to assemble a variety of other data layers for a particular area for some analysis or data exploration, e.g. slope steepness, geologic units, bedding orientation, and vegetation type for a hazard assessment. User centers map view on area of interest, then using an 'add data' tab, accesses a catalog application that allows them to search for web services that provide the desired datasets. After obtaining the results and reviewing the metadata for the located services, user selects one or more to add to the table of contents for the client application. Response from catalog has sufficient information to enable the client application to load and use the resource (e.g. serviceType, OnlineResourceLinkage). More concrete instances of this case would be finding Web Map Services to add as layers in an ESRI ArcMap project, borehole Web Feature Services to post borehole logs in a 3-D mapping application, or water chemistry data Web Feature Service to bring data into a spreadsheet or database.
- User searches for boreholes in an area. Returned metadata records have links to metadata for related resources, like logs of different types, core, water quality data, etc. that the user can follow to browse metadata for these resources.
- A catalog operator wishes to import and cache catalog records from a collaborating catalog that have been inserted or updated during the last month (harvest). This operation requires knowledge of the metadata standard and version used for the returned records.
- A user discovers an error in a metadata record for a resource that they have authored, and wishes to contact the metadata producer to request correction.
- A search returns several results that appear to contain the desired content, and user must select the most likely to meet their needs. Metadata should provide sufficient information to guide this decision.
- A project geologist at Company X is searching for data relevant to a new exploration target, and wishes to restrict the search to resources that are publicly available.
- Complex search examples (see further discussion in the Query complexity section, below):
    - Search based on related resources, for example a search for boreholes that have core.
    - Boreholes that penetrate the Escabrosa formation.

- Sample locations for samples with uranium-lead geochronologic data.
- Find links to pdfs of publications by Harold Drewes on southeast Arizona.
- Find geologic maps at scale < 100,000 in the Iron Mountains.
- Who has a physical copy of USGS publication I-427?

Considerations for the catalog system

Efficient searching

A search should return results that are actually relevant. Existing web search tools are very good at indexing relevance based on association of words in text, and using links and user navigation history for those links. This kind of indexing does not work for datasets, in which the information may be encoded in binary format, and proximity of strings may be a function of the data serialization algorithm, not the semantics. Semantic technology is advancing rapidly, and there is significant effort devoted to increasing search efficiency using background information (common sense) encoded in ontologies. The use of controlled vocabularies (ideally linked to an ontology) to index structured data will enable the system to take advantage of semantic technology to increase search efficiency. Determining the elements requiring such vocabularies must be based on specific use cases.

Identifiers

A widely used identifier scheme is important to reduce duplication, and determine associations between resources. Globally unique identifiers are essential for the described resource, and for the metadata record.

The current thinking in the WWW community appears to be converging on a consensus to use HTTP URIs that are expected to dereference to some useful resource representation. A widely used and understood identifier scheme also enables semantic web functionality. The “anyone can say anything about anything” paradigm requires being able to identify the things.

Query complexity

The complex search examples in the use cases section involve associations between resources, or resource-specific properties. The following table is a decomposition of some complex query examples. Careful consideration of such decomposition is necessary to determine the boundary between catalog services with metadata search, and data services that allow filtering of data elements based on their properties.

Table 4. Analysis of complex queries

Case#	Plain language query	Decomposition	Simplified solution
1	Boreholes that have core in a particular depth interval in a given area.	Borehole-centric approach -- geographic search for borehole resources (assume collar location), filter for those that have a related resource 'core', filter again for property of related resource 'core interval = min, max depth meters'. Alternatively, view search as actually for a 'core' resource, so search should be for 'core' with some given vertical extent. The core resource must provide an ID 'xxxx' for the borehole from which it was obtained. To obtain more details about the borehole, search for metadata on borehole with resource ID = 'xxxx'.	Include keywords for other resources associated with borehole. Put information about these in the abstract. User searches catalog for borehole with keyword (thesaurus=related resource) = 'core', reads abstract to see if it is what they want. The keywords would have to be a controlled vocabulary.
2	Boreholes that penetrate the Escabrosa formation in a given area.	Geographic search for borehole resources (assume collar location), filter for property 'intersects Escabrosa formation'. Alternatively, search for borehole service that includes property = "formation tops", then query that service. Service properties would have to be from controlled vocabulary.	Include names of penetrated formations as keywords on a borehole. Formation names ideally from a geologic unit lexicon.
3	Locations for samples with uranium-lead geochronologic data in a given area.	Search catalog for Geochronology data service with property = 'analysis type' and backtrack to location point through sample metadata, or search catalog for U-Pb Geochronology Data Service and backtrack to location point through sample metadata, or search for 'sample service' with property = 'analysis type'. In the second case, there would still need to be some metadata property to indicate the analysis type for the service. Approach via the analytical data service requires chaining to the sample feature service, analogous to case 1 for borehole service.	Include keywords for kinds of analytical data associated with a sample in the sample metadata record. Search for samples with keyword (thesaurus=analysis type) = 'U-Pb geochronology'.

Case#	Plain language query	Decomposition	Simplified solution
4	Find links to pdfs of publications by Harold Drewes on southeast Arizona.	Search for document resource with author = ‘Harold Drewes’ and geographic extent = ‘SE Arizona’, and online distribution format = ‘pdf’.	Is search by representation format high enough priority to support?
5	Find geologic maps at scale < 100,000 in the Iron Mountains.	Search for geologic map resource with geographic extent = ‘Iron Mountains, and resolution scale denominator < 100000.	Is search by resolution high enough priority to support
6	Who has a physical copy of USGS I-427?	Search for document publisher = USGS, Series ID = I-427, offline distribution format = ‘paper copy’	Include the document ID in the resource description.

Consideration of these queries indicates a requirement to distinguish metadata service from a data service. When the request involves properties of specific instances of a particular resource type, a data service for that resource should be accessed. The metadata for that service should describe the properties offered for resource instances in that service.

Cases 1-3 can be handled in a general way by a service chaining process, in which the catalog is searched for services offering the feature of interest with the property of interest that will be used as a selection criteria. This approach keeps the top level resource catalog simpler, but makes discovery operations significantly more complex. Cases 1-3 can also be handled with scoped keyword terms, where the scope includes things like ‘analysis type’, ‘geologic unit’, ‘related resource type’. In this usage, the scope specifies a controlled vocabulary of categories related to some concept. Addition of new querying capabilities requires adding additional scoped keywords in the metadata. The second approach is viewed as more appropriate in a ‘keep it simple’ design framework for minimum metadata requirements.

Cases 4-6 are related to document-oriented searches, for which distribution format and online access are important, and a number of bibliographic properties (scale, publisher, series, series ID, media, file format) come into play.

Accessing resources

In order for software to utilize URLs in metadata without operator intervention, strong conventions are necessary to guide what URLs are in the metadata and where they are placed. Links in metadata to access resources should in general be complete URL’s that can be invoked with a simple HTTP GET, without having to add additional request parameters. Formal elements (with controlled vocabulary content) should provide machine-processable information to distinguish links that will return a document from links that invoke a service or access an online interactive application. The idea is that sufficient information should be provided that client software can parse the metadata record and provide useful functionality on the resource with minimal user interaction.

For many resources, different representations may be available. These might be different file formats for the same document for information resources. For non-information resources, a variety of representations that have different uses might be available. For example, a physical sample may be represented by a text description of the sample, a GeoSciML xml description, visible light photograph, or images of the sample using other sensors. A geologic map may be available as a paper copy, a scanned image, a georeferenced scanned image, a vector data set in one of several formats (gml, shape file, file geodatabase, MIF, DWG), through a web map service, or through a web feature service. Metadata for a resource should be able to describe all of these different representations that the resource provider wishes to make available, in such a way that automated clients can seek representations useful to that client, or search clients can present users with links to access different formats or representations.

Citation and contact information

Citation information specifies the source of some content. Citations for the described resource specify the origin for the resource intellectual content. The cited agent may have played various roles relative to the resource—author, compiler, editor, collector etc., and a controlled vocabulary is necessary to specify these. Citation for a metadata record specifies the agent responsible for producing the record, typically thought of as the metadata record creator. Metadata production involves elements of authoring, compiling, and editing. Minimally, citations must identify an individual person, an organization, or a role in an organization that is the agent filling a specified role relative to the cited resource. In most cases an organization will be specified, either as the employer or sponsor of a person, an institutional actor, or the host for some role (web master, metadata editor). In addition, information required to contact the cited actor is necessary to enable metadata users to contact a person with some knowledge of the cited resource. For long-lived metadata, contact for an agency role is most likely to persist. The minimum metadata contact information recommended is either an e-mail address or telephone number.

Fitness for purpose

The metadata should provide sufficient description of the resource for a user to determine if the resource is likely to meet their needs, and to determine what representation to access. The simplest approach is to provide such information as text in the metadata abstract, including why the resource was produced, what sort of observation procedures were used, assessment of data completeness, accuracy, and precision, and comparison with other known similar resources. The data quality section of ISO 19115 provides a data structure to formally describe this information, but the cost of using this is high (complex data entry), and there do not currently appear to be clients that utilize the information. The guiding principal should be that if users need to search on some particular quality criteria, specific guidance on how to encode that information in the metadata is necessary (e.g. which ISO 19139 elements, what controlled vocabulary to use if terminology is involved).

Branding

In a distributed, federated catalog system with harvesting, metadata records are expected to propagate far beyond their original point of introduction into the system. If an organization producing metadata wishes to be recognized, and in order for users to be able to contact the metadata originator, contact information for the metadata originator must be considered part of the metadata record, and maintained in harvest processes. For presentation to users, it is desirable to provide a link to an icon that can be displayed with records to brand the origination of the metadata.

The same considerations hold for the resource itself.

Access constraints, legal limitations

Metadata records that are not for public consumption should never be exposed to a harvesting request. Implementation of security and access control must occur at a lower layer in the network stack than the catalog service is operating, such that authorization/ user authentication information is handled by the environment containing the catalog client and server. Metadata for commercially licensed resources may be publicly accessible, but should clearly indicate the licensing requirements and procedure to access the resource.

Low cost of entry

Metadata producers should be able to reuse and build on existing structured metadata. Minimum requirements should be limited to information that is commonly available. Resource specific details should be provided in text elements in the metadata. Special information necessary to utilize web links (e.g. web service operation) in metadata should be provided by text in the metadata or through linked documents.

Requirements Discussion

One of the basic requirements of the NGDS is to make access to data simpler. A major time consuming aspect of bringing disparate datasets together is data integration. This process involves matching field or element names in the schema for various data sets, selecting those that contain the information of interest, and then merging content into a single data set with consistent usage of vocabulary and units of measure in a standardized collection of fields or elements. Data integration in our current system of scientific information interchange is mostly left to the data consumer. One obvious path to simplify data access is to develop standard formats for integrating common data sets (e.g. borehole temperature data, heat flow measurements) that is used to deliver content to data consumers.

A major decision for data delivery in a federated system is where does data integration occur. Until recently, the most common approach was for a data base compiler to collect various datasets and integrate them into a single database that was then made available. This approach works fine while the data compiler has the resources to continue integrating new data, and while the data compiler is still professionally active.

A second approach is to do the data integration as close as possible to the original data provider. By documenting data schema, encoding formats and practices for vocabulary usage, data can be put into the ‘data integration’ format when it is made available on the web. This requires education of the data providers/publishers on the use of the integration formats, but results in a larger community of IT personnel who know how to get data into and out of the integration format. The originator of the data is likely to be a better judge when it comes to making decisions on how to map their content into an interchange format (assuming they are comfortable with the interchange format). In the case of geological surveys or other NGDS data contributors (e.g. subcontractors on other DOE projects), policies can be developed to always present data in the data integration format (along with any other formats that the data publisher wants to use). The net effect is a greater likelihood that the federated information system using the documented interchange formats will outlast any particular researcher, data provider, project, or agency. HTML on HTTP and NetCDF are examples of data integration formats that have achieved wide usage and long term usefulness.

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

System Architecture

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

The system architecture will be described in terms of the functional components shown in Figure 1. These are discussed in the following sections.

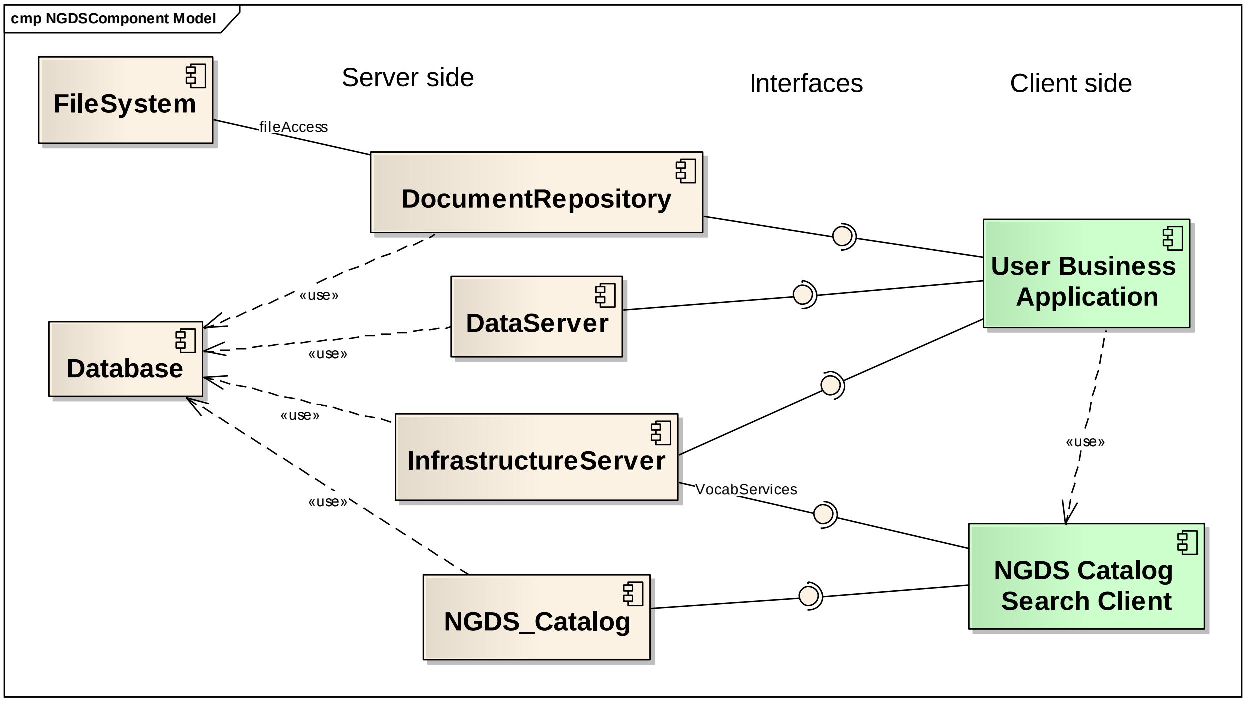


Figure 1. Functional components of National Geothermal Data System. A variety of implementation choices are available for each of the components. Components on the left are mostly hosted by system servers, and interact with the client components on the right through a collection of interfaces defined by the service profiles.

turning metadata. At least one of the implemented protocols and interchange formats used for delivering metadata must conform to an NGDS specification. Initial catalog testing and prototypes are using the Open Geospatial Consortium Catalog Service for the Web (CSW), but other protocols such as the Open Archive Initiative Protocol for Metadata Harvest (OAI-PMH) or the Z39.50 protocol may also prove to be useful. The CSW was selected for initial development work because it operates in the same framework as the other Open Geospatial Consortium services being tested for data delivery (the Web Map Service and Web Feature Service), is designed for geospatial data, and has a variety of free, open-source software projects developing clients and servers for the protocol, as well as a variety of commercial products (including ESRI ArcGIS) that are implementing the protocol.

The CSW service requires all conformant implementations to return metadata using a simple XML encoding of the Dublin Core Elements and Terms, and defines a subset of these that are core queryable and returnable elements (see OGC 07-006r1). The base CSW specification adds a bounding box as a core queryable requirement for any CSW catalog. The CSW service can operate with any xml schema for metadata content, and in the geospatial community, the most widely used profile is for the ISO 19115/19115 metadata. Use of this metadata schema allows richer metadata content that enables greater automation of access to resources.

NGDS Catalog may be implemented with various software and hardware configurations on any node in the system. To be an NGDS compatible/compliant catalog, the only requirement is that they implement an NGDS catalog service profile, and provide metadata in at least one outputFormat schema and profile that conforms to an NGDS metadata interchange specification. Our recommendation is to use the OGC CSW with its base metadata schema (Dublin Core elements and terms), and for more in depth metadata, the ISO 19139 encoding of the ISO19115/119 metadata content model, following recommendations proposed in the USGIN ISO metadata profile.

Document repository

Data in documents will be accessed via URL from document repositories, which are basically web-accessible file systems. In this context, 'document' is used in a very general way as a packaged body of intellectual work with an author (or editor, compiler, or similar originating role), a title, and some status with respect to Review/authority/quality. Documents can be packaged in a single file or a group of related, linked digital files. Documents provide a straightforward path to get data online quickly and easily for the data provider, but if this approach is used for datasets (e.g. Excel spreadsheets, Microsoft Access databases), it requires the data consumer to do all data integration work themselves.

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

Data Servers

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

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

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

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

Thus, in the initial phase of the project services will be defined using simple xml schema with string and numeric-valued elements. These services can be consumed by existing clients like ArcMap and Quantum GIS. These simple schema will be compatible with the ISO specifications to the degree that is practical. As clients are developed for richer-content complex feature services, the NGDS will migrate towards use of the more complex schema. There are also a number of other data formats in use in related communities for geoscience information interchange, including WaterML in use by the CUAHSI project, NetCDF, which is widely used for large numeric data sets in the atmospheric and remote sensing communities, and an xml markup developed for geochemical data by the EarthChem project. Where ever possible, NGDS data providers should reuse existing schema to take advantage of tools developed to consume data in these formats.

Infrastructure Server

The extensive requirements for the NGDS laid out in the requirements section proscribe a collection of functions that must be available on a system wide basis. These functions will be provided by infrastructure servers, prime among which is the NGDS Core at Boise State. The most important infrastructure services that have been identified at this point include caching, mirroring, and backing up system data; providing a home for orphaned data or legacy data; user authentication for access control, vocabulary services for provision of community vocabularies for semantic interoperability, and identifier registration services that will provide URI dereferencing and mapping between identifier schemes to avoid unrecognized duplication of resources. Other infrastructure functionality that would be useful includes validation of information interchange documents to determine if and to what degree they conform to system specifications; and social networking functions such as resource rating, comment, feedback; and usage monitoring and reporting. Development of such infrastructure services should be prioritized to support data services that are actually being implemented.

Database and File System

Various databases and file systems accessed by server applications will house the actual system resources. For security and simplicity, these will probably not be directly accessible for system users, but will be accessed through NGDS servers or clients such as the Geothermal Desktop. Many user applications may also have local data store and file system used to cache resources obtained from the system for offline usage, better performance, and reliability (not dependent on operation of internet).

Clients

The client applications implement most of the desktop analytical and search functionality required by the system. These are outside the scope of this data-access system architecture except for the provision that they operate with the NGDS catalog for resource discovery and evaluation, and utilize NGDS services and repositories for data access.

System deployment

Nodes

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

The deployment diagram indicates a key aspect of the system—the user client software interacts with components on the server side through a pipe labeled “NGDS services.” This connection represents any and all service protocols used to link clients and data servers in the system. These services define inter-

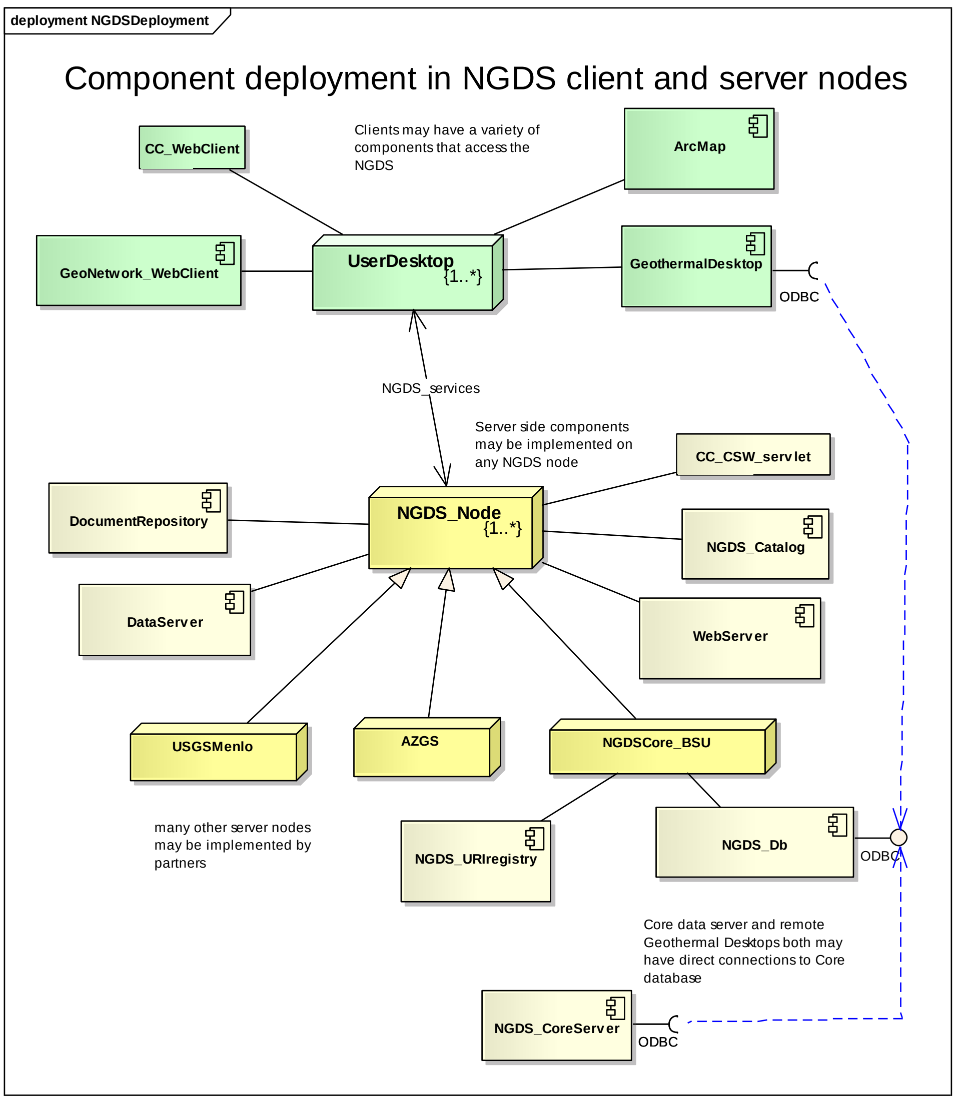


Figure 2. Deployment of components to nodes in the system. The core node will implement special functions, including archives, system specification repositories, and registries of identifiers, as well as standard catalog and data services. Other nodes will implement catalog and data services, and may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

faces that decouple the clients and servers such that as long as the operations and behavior of the service do not change, any upgrades or modifications that occur in the client or server software do not break the system. This loose coupling is a key design feature necessary to allow the system to evolve as technology and user requirements change.

Figure 2 also indicates that direct connections using proprietary technology may exist between clients and servers managed by some participants (ODBC to ODBC connection indicated between client GeothermalDesktop and server NGDS_Db). Such connections may be necessary for expediency, security, or special performance requirements, but should be considered interim solutions because they violate the premise of an open system in which services offered are publicly documented and available to any client in the system.

Data Acquisition Plan

This data acquisition plan is a road map for bringing data into the information infrastructure that is the foundation of the NGDS. This plan is an expanded and fleshed out version of a draft plan that was distributed by AZGS to the NGDS development team in February, 2010. The intention is to get the NGDS off the ground, with useful data content, as quickly as possible by using existing, tested Open Geospatial Consortium services, particularly Web Map Service (WMS) and Web Feature Service (WFS). In a nutshell, the steps in this plan are:

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

Data types for which NGDS data acquisition services and interchange formats have not been specified will be made available in user-defined data files that will be described by metadata in the system catalog and placed in web-accessible servers. Standardization of automated, interoperable data acquisition via services and community interchange formats will be developed incrementally, starting with highest priority data types. Priority will be determined by data availability and requirements from application developers working on the Geothermal Desktop or other client software useful for geothermal resource development.

For interoperable data to be presented to the system using standardized protocols, interchange formats, and vocabularies, the development team will need to work with the user community (data providers and consumers) to determine a useful starting collection of attributes for entities or features that will be delivered, including units of measure and required controlled vocabularies. Interoperability means in practice that software will use the same access protocol for a given kind of information from any NGDS data provider, without any provider specific customization. Some important requirements include:

5. Ensure interoperability among data sets with members adopting common standards and protocols.
6. The data schema must be vetted with stakeholders
7. Data schema for interchange formats must be versioned, such that expanded or modified versions can be introduced without disrupting working systems.

The process of identifying kinds of information to be made available will be pursued on two fronts. NGDS consortium members were polled in January and February, 2010 to get an inventory of the resources that they will be contributing to the system, but the results were limited in terms of specifics, mostly recognizing scanned well logs and other kinds of documents. The data resource inventory has been continuing through verbal interviews by AZGS development staff with information managers at several of the organizations. With the initiation of the AASG geothermal data project, state geological surveys were polled yielding a larger body of data resources to be made available through the system. The evolution of this inventory is continuing as more states develop plans for data contributions, and input from the SMU/Siemens Geothermal Data compilation project is factored in. A current list of data items (resource types, entities…) is presented in Table 5.

Table 5 Summary of data items compiled from AASG data providers (5/27/2010). This listing of data items is being updated and revised based on continuing input from NGDS consortium members, state data providers, and the SMU/Siemens Geothermal data project.

Data Item	Category	Notes for data product
Borehole lithology log dataset	Coverage	Lithology log consists of collections of intervals defined by top and bottom coordinate in borehole trace, and association with lithology description. Lithology descriptions will include original recorded text, lithology categories from CGI vocabulary, and other properties. We anticipate using the GeoSciML schema (http://geosciml .org) for these descriptions, with extensions if necessary. Log will be associated with a borehole collar location, and metadata for the original description. A paper copy of a log or a scan of a paper log is considered a kind of document.
Digital well log	Coverage	A dataset that consists of a collection of measurements of some physical property as a function of depth in a borehole. WITSML (Energistics), NetCDF, LAS are possible interchange formats.
Temperature depth log	Coverage	This is a kind of well log coverage, with the sampling frame corresponding to a borehole, the spatial reference is length measured along the borehole track, measured property is temperature.
Aquifer temperature map	Document	See geologic map. If spatial data are points with temperature measurements, should be considered a
Document	Document	Document is used to mean a packaged unit of content with a single authorship (which may include several people). Examples include books, reports, journal articles, geologic maps, other kinds of maps. Internal content within a document is generally not individually identifiable (unlike records in a dataset).
Geologic map	Document	Geologic maps will be made available through one or more of several mechanism: Download of image file (tiff, Jpg, or pdf), ideally georefer-

Data Item	Category	Notes for data product
		enced; an OGC Web map service based on a map image (no getFeatureInfo) or better on vector data; or as a vector data set. NGDS AASG project personnel can be assigned to assist with implementing map services, but requests for such assistance should be made in the work plan so we can schedule resources. Map services may be hosted by NGDS regional hubs, NGDS core, by the data provider, or any other reliable server, and will be required to be maintained available. Map images as files may be published in repositories
Geothermal map	Document	see geologic map
Gravity map data	Document	See geologic map. Spatial data may be grid or contours.
Resource suitability map	Document	See geologic map. Spatial data could be coverage or polygons.
Active Fault	Feature	A GeoSciML GeologicStructure feature; attributes should include at least a statement of evidence for fault being active; ideally includes orientation information, time since last displacement, hydrologic information about fault zone?
Geologic Unit feature, geothermal characterization	Feature	A geologic unit description specifying properties important for geothermal energy evaluation; includes standard aquifer properties like lithology, permeability, porosity, as well as thermal properties like thermal conductivity and specific heat.
Geologic unit feature, Alteration description	Feature	Is GeoSciML alteration description sufficient, or should it be extended
Geothermal system feature	Feature	Data modeling is still necessary to determine a collection of attributes to characterize a geothermal system as a feature. Subtypes might include developed geothermal systems,
Geothermal system feature, Enhanced	Feature	Needs to be modeled, another subtype of geothermal system?
Hot spring description	Feature	A hot spring is a kind of water source, which is required to have temperature data for the water produced, along with other properties associated with a water source (location, flow rate, water chemistry data, time series for flow and chemistry?). Modeling still necessary.
Intrusive body with heat	Feature	Treat as GeoSciML Geologic feature with geothermal characterization
Sample	Feature	Associated with site; becomes sampling frame for variety of other observations
Volcanic vent feature	Feature	Volcanic vent has a location that may be represented by a point or polygon. Has additional properties that need to be modeled more completely. This will need to be done by partners wishing to contribute this sort of data. A complex data model could be imagined, including eruption history, magma composition, fluid and gas compositions, eruption rates, associated heat flow measurements, associated magma body… For NGDS purposes, we need to identify the key properties of interest as a starting point.
Water source feature	Feature	there are a wide variety of possible characteristics of interest, including chemical, physical, and flow-related properties. Thus a soft-typed approach with {property, measured value} pairs is recommended. We will confer with partners providing this kind of information to establish prac-

Data Item	Category	Notes for data product
		tices for common kinds of characterization. We anticipate being able to use CUASHI practices for much of this kind of data.
Bottom hole temperature	Observation	must have supporting information for the borehole, including location, type of hole (petroleum, mining, groundwater), etc, depth of the measurement, time since between stopping circulation and measurement, the diameter of the borehole at the measurement point.
Chemical analysis (whole-rock chemistry)	Observation	Individual records will be a collection of {measured constituent, abundance pairs}, with identifiers for the analyzed sample, and analysis procedure. The procedure is considered to include the who, how, with what equipment information for the analysis.
Crustal Stress data	Observation	Associated with a site. Needs model
Drill stem test	Observation	An observation feature that includes the results of a drill stem test. Needs modeling of key observation results, including pressure, fluid composition.
Earthquake epicenter	Observation	Treat as observation because epicenter location is always the result of a measurement and analysis process; is observation with result that depends on a collection of seismometer recordings. The epicenter can also be conceived as a feature, with the observation as metadata for definition of the feature.
Flow rate	Observation	Always an attribute of a water source feature or of a water channel feature.
Fluid inclusion data	Observation	Associated with a sample; needs content model
Gravity Station data	Observation	Associated with a site. See PACES for model.
Heat flow measurement	Observation	Heat flow measurements are based on a temperature gradient measured over some interval, and a thermal conductivity value for the material between the two temperature points. The location and temperatures defining the gradient and estimated conductivity must be reported in a complete heat flow report, along with procedure metadata.
Permeability	Observation	May be reported through observation service associating individual samples with permeability measurements. Permeability may also be reported associated with a geologic unit in a
Temperature	Observation	Attributes will include temperature, units, X, Y, Z coordinate, borehole identifier, and measurement procedure. Different measurement procedures will need to be documented. Bottom hole temperature data is one kind of borehole temperature data. Temperature log datasets (as opposed to scanned log documents) are treated as a kind of digital well log. Typically is measured in a borehole to be geothermally interesting.
Thermal conductivity measurement	Observation	Observation feature, attributes include identification of sample used for measurement, procedure, and result with uncertainty.
Trace constituent chemistry dataset. (Water Chemistry, trace-element data)	Observation	Trace element chemical analyses report concentrations of constituents that do not form a significant part of the total material. Most water quality or water chemistry data fall in this category, as well as rock trace-element data. Individual records will be a collection of {measured constituent, concentration pairs}, with identifiers for the analyzed sample, and analysis procedure. The procedure is considered to include the who, how, and with what equipment information for the analysis.
Production sta-		Content will need to be worked out by experts.

Data Item	Category	Notes for data product
tistics record		

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

File based data

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

Data to be scanned

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

Online Digital data

Implementation of online data services will involve several steps. First, an application profile for the service or services to deliver a particular kind of data will have to be developed. In most cases, we anticipate that existing standard services like the Open Geospatial Consortium (OGC) Web Feature, Map, or Coverage services (WFS, WMS, WCS), will provide the necessary framework for services we require. The NGDS technical team will need to develop profile documents specifying the details of how a particular type of information (e.g. borehole temperatures, water chemistry analytical data) will be encoded. Once a profile is in place for a particular data resource, the next step is working with the data providing organization to implement the service with their data.

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

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

For online data services, registration of a dataset in the catalog, and its availability online will constitute 'data acquisition'. Thus, implementation of the catalog as an operational service will need to be one of the first steps in system implementation. AZGS has developed a prototype catalog, implementing the CSW 2.0.2 catalog service using Geonetwork OpenSource, currently at v. 2.6, but in active development with new versions coming out 2-3 times a year.

Technical discussion

Data delivery options

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

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

Data will be considered part of the NGDS when it is locatable using the NGDS core catalog, and accessible via the web according to procedures described in the metadata record obtained from the NGDS core catalog. The anticipated delivery process must be defined in the Statement of Work (Data delivery plan column in spreadsheet) and approved by the project management team before the main data compilation phase of a data development cycle (step 3 in Figure 1, above).

Metadata

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

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

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

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

Summary of Recommendations

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

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

For data services we recommend starting with WFS 1.1.1 simple feature services for a few widely available and geothermally interesting datasets. Based on data compilations thus far, the best candidates appear to be a bottom-hole temperature observation service, heat flow measurement service, and a Quaternary fault service. The content model for xml schema used for data interchange in these services will need to be worked out by the community of data providers as a first step. AZGS has already implemented a demonstration heat flow measurement service using the heat flow data set from the GeoHeat Center.

AZGS is prepared to move ahead with implementation of these recommendations upon approval by the technical advisory group.

References

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

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

ISO 19115:2005, Geographic information - Metadata

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

ISO 19119:2005, Geographic information - Services

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

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

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

USGIN Team, 2010, Use of ISO 19139 xml schema to describe geoscience information resources v1.1, Current version at http://lab.usgin.org/profiles/usgin-iso-19139-profile.

USGIN Team, 2010, Simple Metadata Recommendations for Geoscience Resources, current version available at http://lab.usgin.org/groups/metadata-interest-group/format-agnostic-minimum-metadata-concept.

Appendix T. Comments on “NGDB Repository Metadata for Well Logs” [9-15-2010]

Comments on “NGDB Repository Metadata for Well Logs”

Need to separate metadata for borehole entity from well log entity, and possibly the borehole entity from the collarLocation/well entity (to deal with complex wells that have multiple well-bores). A given borehole may have many logs, various logs will have different extents in bore-hole. Does the metadata for the log reference the metadata for the borehole or duplicate it? Data maintenance/ QA is easier if the log metadata references the borehole.

The borehole itself is a non-information resource (we can’t send the borehole over a wire), but there are metadata associated with that resource that are of interest, particularly what kinds of information resources are associated with the borehole. Logs are a particular kind of resource associated with a borehole; they may have non-information resource representations (a paper copy), or various digital ‘information resource’ representations including a tiff or pdf file containing an image of the log, or LAS, WITSML, or OGC SWE encodings of the actual measured values.

A digital library will maintain resources that are individual file-based representations of various information and non-information resources. From the normal user perspective, in the case of something like the well log, I think people would expect to search for the log in the non-information sense (an observation in OGC terms, independent of any particular representation), and get a description of the observation (where, when, what, who, how), and a listing of the various representations available from the NGDS, including those from digital libraries (repositories), as well as service-based access to the data itself.

Section 1.

“…document is intended to 1) propose an exchange template for well logs into the NGDS Catalog” – Unclear; does this mean a content model for metadata describing well logs in the NGDS catalog?

“Metadata…describes…. Information resources” In the larger scheme, shouldn’t metadata also be useful to describe non-information resources (books, samples, core, interpretations, intellectual works) as well?

Well ID – the authority for the identifier should be specified in some fashion to provide some guidance for dereferencing the identifier.

TypeLog—good idea to specify the basic property measured, but also should include name of log as given by logging company, and possible a name from other log type category schemes

What is syntax for county; state string—‘;’ separator? Is state full state name or abbreviation

Collar Location – if this content model is to be used for constructing queries, have to be more specific

Looks like the assumption is that a well has exactly one wellbore. Somewhere should discuss and justify this choice.

Discussion should have definitions of well and drillhole. I prefer well, borehole, borehole collar, wellbore terminology, as posted in the Data Categories Working group Forum (http://www.geothermaldata.org/NGDSParticipants/DataCategoriesWorkingGroup/tabid/267/forumid/17/threadid/43/scope/posts/Default.aspx)

9a. What is the permanent datum??

Why not report top and bottom of logged interval?

Max recorded temperature—does this need to be reported for interpretation of log trace?

Casing is property of borehole, not of log, but do have to allow for possibility of different logs being run with different casing in place.

Isn’t time since circulation commonly available on a log and useful to know for interpretation

Appendix U. Metadata Crosswalk: BSU Repository – USGIN Recommendations

BSU required and recommended elements are for the NGDS Data Product and Dataset Repository.

BSU Required and recommended elements	***BSU element description***	***USGIN recommendation for discovery/evaluation metadata***	***Comments***
Content Information			
Title	The name given to the Data Product/Dataset by the creator, publisher, or contributing institution.	o **Title** (1 entry): Succinct (preferably <250 characters) name of the resource.	
FileType	The nature or genre of the content of the Data Product/Dataset. For a list of FileType options, see Appendix A.		Generalize to resourceType, cardinality should be 1..N. DC term is 'Type' http://dublincore.org/documents/dcmi-terms/#terms-type
SubjectKeywords or SubjectKeyPhrases or SubjectClassificationCode	Thematic keywords or ClassificationCode describing the topic of the Data Product/Dataset. (N.B. Spatial & temporal keywords are reserved for coverage) or Key phrases describing the topic of content of the Data Product/Dataset.	o Subject Keywords (0 to many entries): Thematic, spatial and temporal free-form subject descriptors for the resource. A keyword may be assigned on metadata import if none are present. If possible, submit keywords in separate Thematic, Spatial, and Temporal keyword categories.	
Description	An account of the content and context of the Data Product/Dataset.	o **Description** (1 entry): Inform the reader about the resource's content as well as its context.	
Source	A reference to a resource from which the present Data Product/Dataset is derived.		a kind of relation? How does this work for a document? Use case I can imagine is bib citation for a map that is the source for digital data set that is the subject of the metadata record. Contrast with LineageStatement?
Relation	A reference to a related resource (e.g., chapter in a book).		dc terms defines several refinements; not very useful unless semantics of relation is specified. To be useful, will need some careful guidance on how the referencing works (is it for people or computers?)

BSU Required and recommended elements	***BSU element description***	***USGIN recommendation for discovery/evaluation metadata***	***Comments***
CoverageTemporalKeywords	The temporal period (date, date range, a named period label) of the content of the Data Product/Dataset.		is this a keyword or a date/date range?
		o Temporal Extent – Temporal range over which the resource was collected or is valid. If the resource pertains to specific named geologic time periods, those terms should be entered as keywords (preferable as part of Temporal Keywords). Start Date (0 to 1 entry), End Date (0 to 1 entry; required if start date exists),use ISO 8601 date and time format.	
CoverageGeoSpatialKeywords	The spatial location (place name) of the content of the Data Product/Dataset.		
Intellectual Property			
ContributingInstitution	The organization(s) responsible for making the Data Product/Dataset available.		I think this is probably equivalent to the DistributionContact in AZGS recommendation.
Publisher	The service responsible for making the Data Product/Dataset available.		Won't this be in the BibliographicCitation. What is the use case that requires having a separate publisher field? To be useful, there would need to be a registry/controlled vocabulary of publishers to get around the variety of abbrevations that are used in publisher names...
Creator	An entity primarily responsible for making the intellectual content of the Data Product/Dataset (author).	o **Originators** (1 to many entries): Authors, editors, or corporate authors/curators of the resource.	We have had quite a time distinguishing creators from contributors…
Contributor	An entity responsible for making contributions to the content of the Data Product/Dataset (co-authors/organizations).	o Originators (1 to many entries): Authors, editors, or corporate authors/curators of the resource.	
Rights	information about rights held in and over the Data Product/Dataset (a link		need clarification on use of this element as opposed to ContraintStatement. Are both re-

BSU Required and recom-mended ele-ments	***BSU element description***	***USGIN recommendation for discov-ery/evaluation metadata***	***Comments***
	to a copyright notice).		quired?
Instantiation (Version)			
Language	A language of the intellectual content of the Data Product/Dataset (defined by 2-3 letter primary language tags)	o Resource Language (0 to 1 entry): Use three letter ISO 639-2 language code (defaults to "eng" for English).	
DatePublished	The publication or release date of the Data Product/Dataset.	o **Publication Date** (1 entry): Publication, orig-ination, or update date (not temporal extent) for the resource. Use a "year" or ISO 8601 date and time format. Alternative date formatting must be machine readable and consistent across all datasets. If no publication date is known, estimate the publication date range, enter the oldest year as the publication date, and include the estimated date range in the Description field	
Format	The physical or digital manifestation of the Data Product/Dataset type. This field can include the dimensions of the Data Product/Dataset.	o Distribution Keywords (0 to many entries): keywords describing the physical form of the resource (core, rock sample, digital file, book, journal article), formatting of resource content (file format, e.g. tiff, xls, MIME type), or physi-cal distribution media (film, floppy disk, online service, hard copy). Table 6 in USGIN ISO metadata profile includes a vocabulary for dis-tribution format for use with the ISO19115 dis-tributionFormat name property. Use of these keywords allows users to search for particular kinds of artifacts	
Identifier	An unambiguous reference to the Data Product/Dataset within a given context (e.g., ISBN/DOI/local ID-from publish-er).	o Resource ID (0 to many entries): Resource identifier(s) following any public or institution-al standard. Identified consists of an identifier string and if applicable a Resource ID Protocol identifier string that specifies the protocol for the resource ID standard. For example: unde-fined, ISBN-10, ISBN-13, ISSN, URN, URI, IRI,	

BSU Required and recommended elements	***BSU element description***	***USGIN recommendation for discovery/evaluation metadata***	***Comments***
		DOI, HTTP, SSN, etc. Examples: doi:10.1000/182; isbn:0-671-62964-6; issn:1935-6862; azgs:OFR-10-02. Many protocols build the identifier for the protocol into the identifier string.	
Bibliographic-Citation	The recommended reference to be used for the published Data Product/Dataset.	o Bibliographic Citation (0 to 1 entry): Full bibliographic citation if the resource has been published.	[AZGS]: 2. Published documents require a standard bibliographic citation (author, year, publisher, series, volume, page numbers, etc.) as specified by a publication style or guideline. Some example guidelines include USGS Suggestions to Authors and MLA Style Manual; the community will need to agree on conventions to use for citation syntax to improve interoperability. In general, for web-accessible digital resources that are the typical items of interest that will be cited, full text searches are anticipated to be the most common use case. Unless clear examples of use cases requiring more disaggregated representation of citations in the metadata (e.g. separate attributes for publisher, larger work title, larger work editor, volume, issue number, etc…) we will stick to simple text blob citations.
Optional fields to be populated for optimal metadata accessibility.			
Coverage Discrete Geospatial Data (FGDC)¤		o Geographic Extent - Horizontal (1 entry, point or minimum bounding rectangle): North Bounding Latitude, South Bounding or Point Latitude, East Bounding Longitude, West Bounding or Point Longitude. Values given in decimal degrees using the WGS 84 datum. *Some resources may not be usefully described by an extent; if no extent is specified the default is Earth.* This convention would have to be modified for systems describing extrater-	

BSU Required and recom-mended ele-ments	***BSU element description***	***USGIN recommendation for discov-ery/evaluation metadata***	***Comments***
		restrial resources. If a particular encoding scheme requires a bounding box, a minimum bounding rectangle will be created if only a point coordinates is given.	
Latitude (sin-gle point data)	" -90.0<=G-ring Latitude <= 90.0 "	see above	
Longitude (sin-gle point data)	" -180.0<=G-ring Longitude < 180 "	see above	
NorthBound-ingLatitude	Northern-most coordinate of the limit of coverage expressed in latitude; -90.0<=North bounding Coordinate <= 90.0	see above	
SouthBound-ingLatitude	Southern-most coordinate of the limit of coverage expressed in latitude; -90.0<=South bounding Coordinate < 90.0	see above	
WestBound-ingLonditude	Western -most coordinate of the limit of coverage expressed in longitude; -180.0<=West bounding Coordinate < 180.0	see above	
EastBound-ingLonditude	Eastern -most coordinate of the limit sof coverage expressed in longitude; -180.0<=East bounding Coordinate <= 180.0	see above	
Units	Decimal degrees, Decimal minutes, Decimal seconds, Degrees and decimal minutes, Degrees_ minutes_ and dec-imal seconds, Radians, Grads, meters, US feet, international feet	see above	
GCDa-tumName	Name of the Geographic Coordinate Datum	see above	
MapProjec-tionName	Name of the mathematical transfor-mation of a 3D surface to a flat/2D	see above	

BSU Required and recommended elements	***BSU element description***	***USGIN recommendation for discovery/evaluation metadata***	***Comments***
	map surface (e.g., Mercator)		
Cell Size (applicable for Raster)	Raster grid cell size; this field is mandatory when the Data Product/Dataset is a Raster		[AZGS]: 3. Spatial data specification require information on spatial resolution and terms to categorize spatial representation type: raster (spatial array), polygon, lines, and points. For maps typically want to record scale to indicate resolution.
		o Geographic Extent – Vertical (0 to 1 entry*): Datum Elevation, Datum Type, Maximum Elevation, Minimum Elevation. Values given in meters. Maximum and Minimum Elevations are relative to the reported datum elevation, which will typically be the Earth surface at the location of the resource or sea level. Datum Elevation must be reported relative to mean sea level (MSL) in meters using EPSG::5714 geodetic parameters (WGS 84). Datum type must be a controlled vocabulary (Earth surface, MSL, Kelly bushing, etc.). The maximum is always numerically greater than the minimum elevation. For boreholes with datum at the earth surface, depth below surface is reported as a negative number. *Vertical extent may be reported relative to different datum (e.g. sea level, Earth surface) in the same record. Example: core from borehole at depths between 100 and 470 feet, borehole collar at 4787 feet above sea level. Vertical extent could be reported in either of the following ways: {0, “MSL”, 1420, 1308} or {1450.6, “Earth surface”, -30.3, -142.4}	[AZGS]: Vertical extent is required for resources that pertain to a subsurface, ocean, or atmosphere location. If no vertical extent is specified, it is assumed to be the current Earth surface.
Source (USGIN)			
LinktoSource	A URL pointing to a resource or resource webpage	o Link to the resource (0 to many entries): A URL pointing to a resource or resource	

BSU Required and recommended elements	***BSU element description***	***USGIN recommendation for discovery/evaluation metadata***	***Comments***
		webpage. URL, Link Function, Representation Format. URL is minimum content required if a link is included. Optionally, a Link Function term from the ISO19115 OnlineFunctionCode controlled vocabulary specifies what a HTTP GET using the URL will invoke. The link might return an html page, electronic document in some other format, an end point for a service, an online application that requires user interaction, etc. Representation Format is a controlled vocabulary term specifying the format (MIME media types) of a file-based response if applicable	
AccessInstructions	A sentence or paragraph describing how to access the information	o **Access Statement** (1 entry): Text instructions for how to access the resource	
QualityStatement	Describe the quality of the Data Product/Dataset	o Quality Statement (0 to 1 entry): Text specification of the quality of the resource	
ConstraintsStatement	Describe the Data Product/Dataset's legal and usage constraints	o Constraints Statement (0 to 1 entry): describe the resource's legal and usage constraints	
LineageStatement	Describe the Data Product/Dataset's provenance	o Lineage Statement (0 to 1 entry): Text description of the resource's provenance	
Contact Information (Creator; Contributor; Cataloguer; Contributing Institution; Publisher)		o Contact - Author or Intellectual Originator (0 to 1 entry): The primary party responsible for creating the resource. Organization Name, Person Name, Street Address, city, State, ZIP Code, Email, Phone, Fax, URL. If contact information is provided, include at least the organization or author name.	
		o Distribution Contact (1 entry): The party to contact about accessing the resource. **Organization Name or Person Name,** Street Address, City, State, ZIP Code, **Email,** Phone, Fax, URL.	have to consider what is rational for separate fields for address, city, state, zipcode, country on contact information? Will anyone every search for resources whose contact is on "Rain-

BSU Required and recommended elements	***BSU element description***	***USGIN recommendation for discovery/evaluation metadata***	***Comments***
		In general, a contact for distribution should be required for physical resources. The USGIN recommendation for simplicity is to require at least an organization or person name and an e-mail address for contact.	bow Ridge Road"? I think our requirements are met by simply including a postal address text field for all of this, which can be searched using a free text search to address any use case I can think of...
Name	First and last name of creator/contributor/contributing Institution/ publisher/cataloguer	see above	
Address	Physical mailing street address	see above	
City	City of address	see above	
State	State/territory/administrative district/province of the address	see above	
Zipcode	Postal code unique to the city of address	see above	
Country	Country of address	see above	
Email	Electronic mail address	see above	
Phone	Telephone number by which individuals can speak to the organization/individual	see above	
Fax	Fax number	see above	
URL	Uniform Resource Locator used to access the contact's website (if available).	see above	
Additional Information	String/Text field; All other pertinent information on the metadata may be entered here.		This should be included in the description field; what is gained by putting it in a separate field?
		o **Metadata Date** (1 entry): Last metadata update/creation date-time stamp in ISO 8601 date and time format. This may be automatically updated on metadata import if a metadata format conversion is necessary.	this and subsequent fields are necessary in a distributed, harvesting catalog system.
		o Metadata Contact (1 entry): The party to contact with questions about the metadata it-	

BSU Required and recom-mended ele-ments	***BSU element description***	***USGIN recommendation for discov-ery/evaluation metadata***	***Comments***
		self. **Organization Name or Person Name,** Street Address, City, State, ZIP Code, Email, Phone, Fax, URL.	
		o **Metadata Specification** (1 entry): Identifier string for the metadata specification used to create a metadata record encoding this con-tent. Should indicate the base standard and version, as well as any profile that applies to the content or encoding. Ideally the identifier could be dereferenced to obtain information about the applicable specification. Identifiers for metadata encoding specifications to be used in the USGIN and NGDS systems will have to be formally defined and registered for such identifiers to be broadly useful.	
		o Metadata UUID (0 to 1 entry): A Universally Unique Identifier (UUID) will be assigned dur-ing the metadata import process if one is not provided. Unique identification of each metadata record is required to avoid duplicate entries across multiple metadata catalogs. The UUID format provides unique identification without centralized coordination	
			4. Web Services require: a. service type from controlled vocabulary (this could be resource type [aka FileType] in this medata scheme). See Table 11 in USGIN ISO metadata profile for a starting-point interim vo-cabulary. b. URL for service-specific document that de-scribes operation of service (e.g. OGC GetCapa-bilities, WSDL). could be done through related, but need relationship type term on relation el-ement.

BSU Required and recom-mended ele-ments	***BSU element description***	***USGIN recommendation for discov-ery/evaluation metadata***	***Comments***
			c. Base URL for service requests. This could be the resource URL supplied. d. Contact information for service provider. For service use distributionContact.
			Entity and attribute data for datasets and ser-vices need to be accounted for to guide people searching for datasets with particular kinds of information. One possible solution is a high lev-el data item/entity/feature vocabulary and a property vocabulary. This would require metadat creators to map the actual entity and property names in their datasets into the con-cepts in the vocabulary. Probably a useful in-teroperability/data integration exercise anyway, but might be non -trivial... Another approach might be to embed FGDC style or ISO19110 en-tity-attribute elements in the DC xml. This would probably help someone trying to use the dataset, but wouldn't provide reliable search results if the entity and attribute names aren't standardized.