METADATA

FOR

GEOSCIENCE INFORMATION RESOURCES

CONTRIBUTIONS TO THE NATIONAL GEOTHERMAL
DATA SYSTEM

Notes from webinar held Feb. 16. 2011

Speaker Stephen M. Richard

Compiled and edited by Leahanna M. Musil

Arizona Geological Survey

A brief overview on how to use the catalog to create metadata records for resources. There a handful of catalog applications to choose from when creating metadata. The catalog applications are very similar, yet may call for slight variations in metadata content needed to validate a metadata record. First you must know what resources you are going to put into the catalog. When creating many metadata records, all belonging to a similar theme (area, publications in order from year, authors, Geological Era, minerals, etc...) it's best to create a collection so they can be easily searched by the public (if for public use) and then can be easily updated and edited by the steward. After you have your resource, a collection idea, and the catalog application you want to use; you are ready to start creating metadata records.

CONTENTS

● INTRODUCTION

● Catalog

The catalog is a collection of metadata records describing resources available in the information network.

● Metadata Records

You will need a resource before you can make a metadata record; a resource can be anything from PDF files, papers, special publications, physical samples, and especially information that might not exist anywhere else.

Structured. The metadata records within a catalog will be structured with consistent information, arranged to enable automated extraction/location of particular information.

Shared. Records are also shared. Each record 'advertises' a resource. They are meant to be distributed widely and easily accessible.

Managed. Metadata records are owned by a 'steward' who is responsible for creating and maintaining the metadata record. The steward is expected to keep metadata content accurate and up to date, and to correct errors that are reported.

Single Metadata record for each resource

The objective of the catalog system is to contain one metadata record for each available resource. In order to realize this goal, project partners create and submit record for resources they have created or that are unique to their holdings, which they expect no one else will be claiming. In the long run the idea is to develop processes to comb the catalog system looking for and reducing duplication. For now the plan is to be aware of duplication by not creating metadata for resources that may have already been made available by another resource steward. Resources like publications, information coming from the Geothermal Resource Council, OSTI publications, many DOE publications, and data from the GEOHEAT Center or EGI that are being handled under the offices of the BSU project, are all things to be aware of that might have metadata already created for them.

Data versus Metadata

Metadata is used to find an identifiable item of interest, a resource, which is available electronically or physically. Typical search criteria include a bounding box, keywords, author, title, dates, etc. The intention of the catalog is to describe what might be called 'data products'—packaged units of information that would be delivered as a document or an electronic file. Individual data items or records are not expected to have metadata in the catalog, although they may very well have more granular item-level metadata that is delivered through services specific to that data type.

Metadata records in catalog. Examples of resources with metadata records would include documents, physical samples, datasets, services, software, files, and databases

Metadata records not in catalog. Examples of resources generally not included in the catalog would include those with domain specific services, such as individual chemical analyses, temperature observations, gravity stations, strike-and-dip measurements, or rock descriptions.

The reason for this is that when you search for these items, you may want to know specific information, e.g. the concentration of silica present or whether gold was detected in the sample. These items would need a more domain-specific delivery format and query protocol than could be implemented for the general, cross-domain catalog.

● Resource

A resource is any identifiable item of interest.

● Collection

For data entry purposes, metadata records may be grouped into collections of records sharing some subset of properties, e.g. same distributor or metadata creator.

● METADATA CONTENT

The purpose of the metadata is to enable people to find useful resources efficiently. A resource in this context is any identifiable item of interest. Most of the resources of interest for the Geothermal Data System will be electronic and can be delivered via the web. The intention is that people can easily discover (find), evaluate, and obtain the resource quickly.

The basic content model for the metadata follows that outlined by the Dublin Core Metadata standard (http://dublincore.org/documents/dces/). The content is mostly specified by brief text describing the resource. It is important to supply as much information as possible. Especially

when you're dealing with the resource information (Title, description-abstract, Author, etc.) you want the person viewing the metadata record to know what they have found and understand what it contains and how the content is represented.

The following content model follows the USGIN recommendation (http://lab.usgin.org/profiles/doc/metadata-content-recommendations). For detailed discussion of metadata encoding and content issues, see http://lab.usgin.org/usgin-iso-metadata-v1-1.

- **Title:** name of the resource. Titles in metadata records should attempt to uniquely identify the described resource for a human reader—in many search applications, the first search results returned will be a list of titles, so these should be as informative as is reasonable in one or two lines of text.

- **Description:** Inform the reader about the resource's content as well as its context. This will be one of the most useful metadata sections for human users to understand the resource. If the resource has an abstract, it is useful to copy that here. Any other details about content formatting, data acquisition, quality considerations, or access constraints that would be important to someone evaluating the resource for their application is also useful. Many searches are done using simple text indexing of the content of the title and description, so a good description will improve search efficiency.

- **Originator(s)**: Authors, editors, or corporate authors/curators of the resource. In order to make searching more efficient, consistent formatting of personal names is highly recommended, with first and last names fully spelled out. Recommended convention is 'LastName, FirstName MI'.

- **Publication Date**: Publication, origination, or update date (not temporal extent). If no publication date is known, recommended practice is to provide your best estimate. If the publication date is a date range, enter the oldest year as the publication date, and include

the estimated date range in the Description field. This is important for searching purposes, when people are looking for a specific resource.

- **Geographic Extent - Horizontal** (point or minimum bounding rectangle): North Bounding Latitude, South Bounding, East Bounding Longitude, West Bounding Longitude. Because some catalog search software malfunction if the corner points of the bounding box are coincident, point locations will be reported at the SW corner of the bounding box, and the NE corner will be offset very slightly to create a non-degenerate box in the metadata record. Values are given in decimal degrees, and the recommended spatial reference system is WGS 84. 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 extraterrestrial resources.

- **Contact Information for the Author or Intellectual Originator**: The primary party responsible for creating the resource: Organization name, Street Address, City, State, ZIP Code, Email, Phone, Fax, and URL. The minimum requirement is either an e-mail address or voice telephone number.

- **Bibliographic Citation**: Full bibliographic citation if the resource has been published. This should include author, title, date published, publisher, other standard bibliographic information following standard journal citation conventions. This is the recommended citation for the described resource for use in publications utilizing the resource.

- **Subject Keywords**: Free-form subject descriptors for the resource. A keyword may be assigned on metadata import if none are present. When using a template the keywords you will need to use the pipe character ("|") to separate more than one value. When parsed, the values from keywords will be put into the metadata separately.

    Thematic: Theme

Spatial: Place

Temporal: Time interval – Geologic Era (Paleozoic, Jurassic, etc.)

- **Resource Language**: The resource language will not always be in English. This is useful information for those seeking a resource in their own language. Three letters are used to identify the language. Default set to eng' for English.

- **Resource ID**: The Resource ID is an identifier string that is unique to the resource and, if applicable, a Resource ID Protocol identifier string that specifies the protocol for the resource ID standard.

   Examples: **doi:10.1000/182; isbn:0-671-62964-6; issn:1935-6862; azgs:OFR-10-02.**

- **Geographic Extent – Vertical**: Datum Elevation, Datum Type, Maximum Elevation, Minimum Elevation. Values given in meters. Maximum and Minimum Elevations are relative to the reported datum elevation. If the datum elevation is 0, the vertical extent is assumed to be an elevation, reported positive upward. If the datum elevation is non-zero, the vertical extent is assumed to be a depth, reported positive down from the datum elevation. For boreholes that are collared at precisely sea level, the elevation convention (downward is negative) must be used for consistency. Datum elevation must be reported relative to mean sea level (MSL) in meters; this convention is necessary for metadata interoperability. Datum type must be a controlled vocabulary (Earth surface, MSL, Kelly bushing, etc.). The vertical extent maximum is always physically above the reported vertical extent minimum elevation.

- **Quality Statement:** Text specification of the quality of the resource.

- **Lineage Statement:** Text description of the resource's provenance. A narrative description of how resource originated and what kind of processing occurred to get it into its current state.

- **Access Statement:** Text instructions for how to access the resource.

- **Distribution Contact**: The party to contact about accessing the resource. Organization Name or Person Name and e-mail or voice phone number are required. Street Address, City, State, ZIP Code, Fax phone number and URL are optional. A contact for the distribution agent is required for physical resources.

- **Link to the resource:** A URL pointing to a resource or resource webpage. The link might return an html page, electronic document in some other format, an end point for a service, or an online application that requires user interaction. A complete metadata record should include information specifying what the link does and the format of the representation retrieved by the URL.

- **Constraints Statement**: Text statement specifying any legal or security constraints on access to or use of the described resource. Answers questions like 'Who can use this resource?', or 'Is it public?'

- **Distribution Keywords**: Keywords describe 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 physical distribution media (film, floppy disk, online service, hard copy). Table 6 in USGIN ISO metadata profile includes a vocabulary for distribution format for use with the ISO19115 distribution Format name property. Use of these keywords allows users to search for particular kinds of artifacts.

- **Metadata Date**: Last metadata update/creation date-time stamp in ISO 8601 date and time format, which looks like 'yyyy-mm-ddThh-mm'. This may be updated automatically on metadata import if a metadata format conversion is necessary.

- **Metadata Contact:** The party to contact with questions about the metadata itself. A person, position, or organization name and e-mail or voice phone number are required. Street Address, City, State, ZIP Code, Fax phone number, and URL may be included.

- **Metadata Specification**: Identifier string for the metadata specification used to create a metadata record encoding this content. 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. The USGIN profile mandates "ISO-USGIN" for the metadata standard name.

- **Metadata UUID**: A string that uniquely identifies the metadata record. 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. In order to avoid multiple identifiers for the same metadata record, the metadata owner should assign a UUID when the record is created, but a UUID will be assigned by system catalogs during the metadata import process if one is not provided.

* For more detailed information on the metadata content model use the following link.

http://lab.usgin.org/profiles/doc/metadata-content-recommendations

● SEARCHING AND VIEWING METADATA IN THE CATALOG

Most applications for searching metadata will be similar, with only slight differences. In this example we will be using the USGIN Geoportal site.

http://catalog.usgin.org/geoportal/catalog/main/home.page

After clicking the link above you will arrive at the home page for the USGIN Geoportal. To start searching metadata records you can either type in the desired search criteria on the home page search box, or you may click the tab next to HOME that's says "search."

*The actual search page will give you more options to narrow down your search.

After typing in the search criteria you choose, click search and items matching your search will be on the right hand side of the page.

Metadata for Geoscience Information Resources in the National Geothermal Data System   9
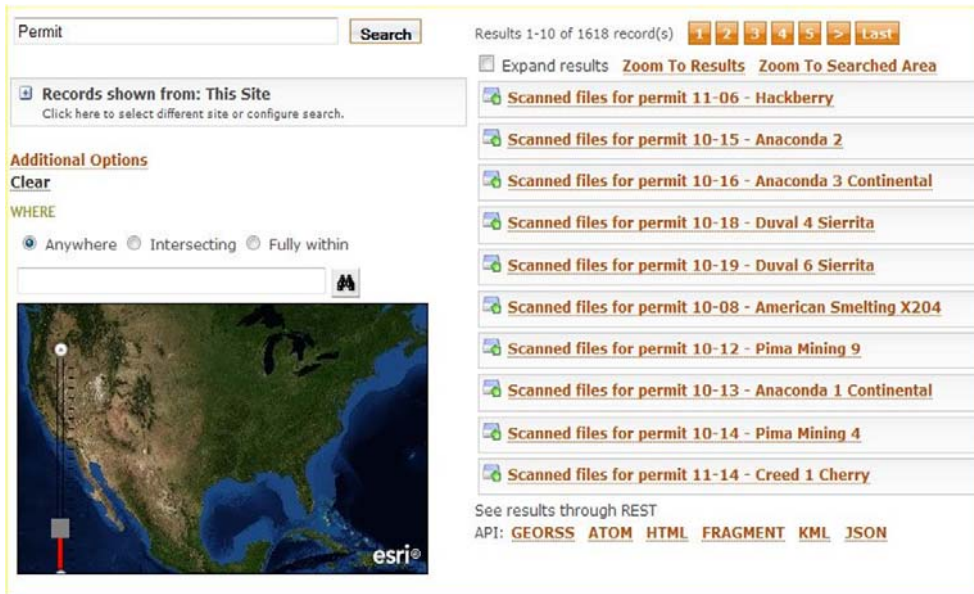
After skimming through the items, find one that you wish to view and then click on the title for the metadata record. After doing so, a row of options will appear: Open, Preview, Details, Metadata, and Zoom To. These are all ways to view the metadata record.

> ➤ By clicking Open, the resource for the current metadata record will open for you to view.

➤ By clicking preview you will receive a preview to either the location on a map, or the resource associated with the metadata record.

➤ By clicking details, you will be taken to the metadata details page. This will include the metadata content.

➤ By clicking metadata, you will be reviewing the XML for the metadata record.

➤ By clicking Zoom to, you will zoom in on the map, set for the location of resource.

HOW TO CREATE A METADATA RECORD

There are two main work flow situations outlined here. First is the situation in which you already have metadata in some kind of structured form that can be manipulated into the metadata template table structure. Second is the situation where the metadata is being created from scratch, in which case online data entry tools are available for use.

Data vs. Metadata when creating metadata records

The distinction between data and metadata is based on how much about the resource and its content you can successfully describe. For the purposes of the catalog, metadata is intended for cross-domain resource description intended to aid discovery and evaluation, but not data analysis. If you find it hard to describe some information in the metadata content model, it's an indication that you should be delivering that information using a data service, not the catalog.

Metadata in a Structured Form (Excel, CVS, Microsoft Access, SQL, Python)

The main idea is to have metadata content in tables that can be extracted, by loading the data from any kind of structured format you have into the template file. Right now what we have online at the USGIN Document Repository (http://repository.usgin.org/uri_gin/usgin/dlio/162) are templates that simply define a table structure with header fields and content listed.
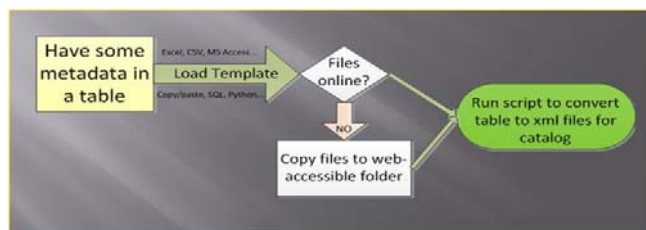
There are all kinds of possible work flows for getting information from one structured format, whether it's a table or XML file, into another structured format (Microsoft access, excel, or another relational database) that allows you to set SQL queries. You can use python scripts to do some of the work for you. Even with Excel you can do a lot just by going through data, copying and pasting.

The "black box" in the scheme is the trick to get your content into the template, whether it's going to represent that template in an access form, a TriSpectives 3-D drawing (TSB) file, or Excel spreadsheet.

Once in the template, you want to make sure that the resource you have and want to make accessible is available for online access either through your service or another service. If already online, you copy the URL for the resource file location and enter it into the template URL section.

If you don't have the files online already, you're going to have to figure out where you're going to put them. You can either put the files in a web accessible directory on a server you have or they can go to one of the hubs.

After that is done, we a have script that runs against the table data turning it into a text file. Then another program is run against that and creates XML files that we then put into a web accessible directory which points to the Geoportal, our catalog that harvests them, and the record ends up in the metadata catalog to be searched.

This is a Web-accessible Directory containing links for multiple Web-accessible folders (for different collections, catalogs, resource type etc...)

This directory has all the metadata which were created by processing a table. We point the Geoportal at this directory and it reads these files and created a metadata record in the catalog.

The workflow steps are:

 ➢ you enter the metadata into the template;
 ➢ you make sure you also have the URL's (that point to the file location online) in the template; and
 ➢ take that script and run it or make that script available online so anyone can run it.

Just make sure to have the URL's pointing at your own web accessible directory.

This is a Web-accessible Folder containing the metadata files in XML format.

Metadata content should match the template file. To download the template file use the link below then follow the page down for the Zip archive, containing MS Excel Workbook with template. Viewing this will also give you an example of the content inside the template.

http://repository.usgin.org/uri_gin/usgin/dlio/162)

> ➢ You'll need to use the pipe character ("|") with keywords when you have more than one value to separate them. When parsed, the values from keywords will be put into the metadata separately.

> ➢ Sections highlighted in pink indicate what is required in the template.

> ➢ When importing and exporting data from one file to another, run through and make sure data wasn't altered in the process. Templates can also give you more control not duplicating material.

● Starting from Scratch

When starting from scratch, meaning you have no metadata set up in some structured form, such as an excel spreadsheet, you will establish first if you have the resources you would like to make records for online access.

If the resource has a file online already, you will be using the Metadata Wizard.

If you need to upload your resource file, you will be using the Document Repository.
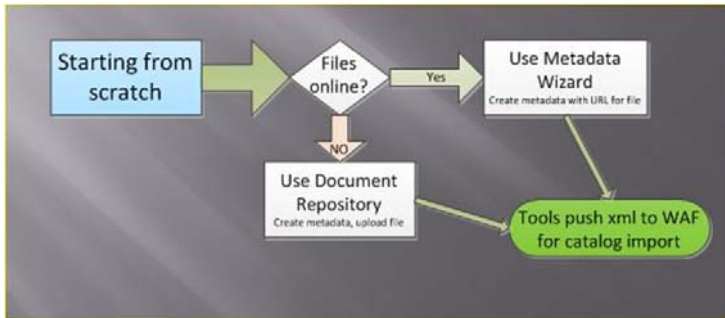
Metadata Wizard
http://mw.usgin.org/
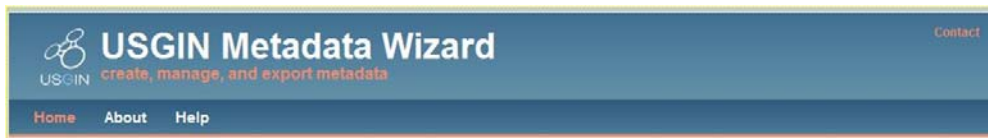
After arriving to The Metadata Wizard home page you will see a flow chart that briefly explains the tools of the Metadata Wizard (create metadata with URL for file). For a more detailed description of each box in the flowchart, point to for a pop-up box for a detailed description for that tool.

To get started in the Metadata Wizard, you will need to sign in. There is a box that includes six email options, located on the left hand side of the home page. This includes Google, Yahoo, Windows Live, OpenID, Blogger, and AOL/AIM. If you do not have a profile with any of these, you will need to create one.

After signing in you will get started by clicking the DASHBOARD option, located on the top left hand side of the Home page. Here in your DASHBOARD, you will see your collections. You may create a new collection or join an existing collection.

Dashboard

Metadata Administration Dashboard (MAD)

My Collections   My Submissions   All Submissions

Curated Collections

Add a new Collection   or   Join an existing Collection

If you plan on joining an existing collection, after you click the link you will be brought to a page that you can scroll up or down on to choose which collection to join.

After choosing a collection, there will be a 'request membership' link on the right hand side of the collection name. After sending your request, the Manager of that collection will need to provide approval for you to view, add and update metadata in this collection.

If you prefer to create a collection or your own, return to dashboard and click the 'Add a new Collection' link on the top left of the dashboard page. After clicking the link, you will be brought to the Create Collection page. You will text boxes for you to enter information about your new Collection.

Now return to your dashboard, you should have more options available. You may view post, 'View Member', 'Export', 'View Summary', 'Add Templates' and 'Add Content'. Click 'Add Content' to create metadata.

Groundwater Quality Reports

Created by: lmusil27
Most Recent Revision: Tue, Nov 16th, 2010 -- 4:10 pm by Diane

Export: Export Help.
CSW Insert XML, CSV File
Summary:
HTML Report, DOC Report

Add Content
Add Template
Edit Collection

Collection Management
2 members
add or remove members
21 posts

Templates
No templates available.

You are now on the Create Metadata page; this is where you enter information. There are eleven labeled tabs on the left hand side. Click on any tab to enter information that corresponds with the metadata record.

(* A red asterisk indicates data that is required.)

When you have finished filling in your metadata information, make sure all required information is provided. You may preview your metadata, or you can save it. You may also publish your information (if you are not the manager of the collection, you will have to wait for it be approved by the manager).

At the top of the page, point to the dashboard link, and click 'My Submissions'. Here you will be able to see everything you have submitted. You can click on a metadata record to View, Edit, Track, or Clone it.

For more Information, click the Link bellow for a detailed Tutorial of The Metadata Wizard. (http://lab.usgin.org/sites/default/files/application/file/u1/USGIN_Metadata_Wizard_Tutorial.pdf)

USGIN Document Repository
http://repository.usgin.org/

In the USGIN Document Repository, you can create metadata and upload resource file. After arriving at the USGIN Document Repository, you will need to login/ or create an account. Create an account by clicking the Register button on the top right hand side of the home page. (* A red asterisk indicates data that is required.)

Note: In the 'Affiliation' and 'About You' section, please be as detailed as possible so we can ascertain you are using the USGIN Repository for the appropriate reasons.

After logging in to the USGIN Document Repository, you should be directed to your DASHBOARD. In your DASHBOARD, you will see your collections. You can create a new collection or join an existing collection.

If you plan on joining an existing collection, after you click on the link, you will be brought to a page where you can scroll up or down on to choose which collection to join. After choosing a collection, there will be a 'request membership' link on the right hand side of the collection name. After sending a request, the Manager of that collection needs to approve you for you to View, Add, and Update metadata in this collection.
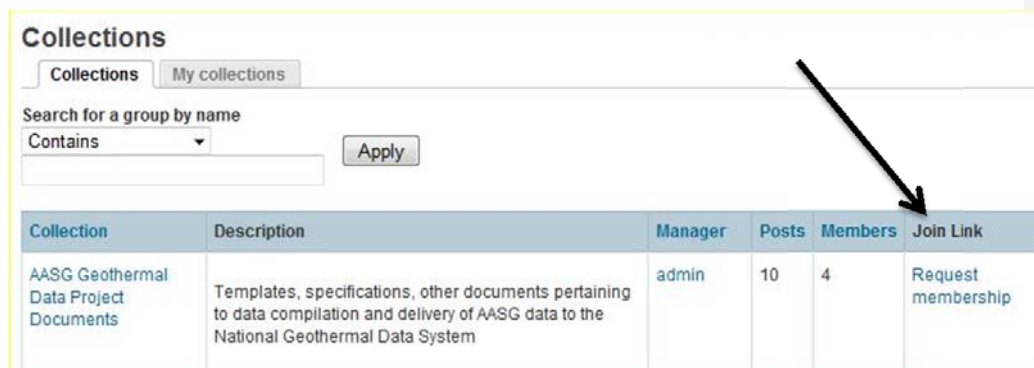

If you prefer to create a collection on your own, return to DASHBOARD and click on the 'Add a New Collection' link on the top left of the DASHBOARD page. After clicking the link, you will be brought to the Create Collection page. You will now see text boxes for you to enter information about your new Collection.

Return to your DASHBOARD. Now you should have more options available. You may view post, view member, invite members, see published posts for that collection, subscribe to notifications, edit subscriptions, add templates, and add content. Click on Add Content to start creating a metadata record for your collection.
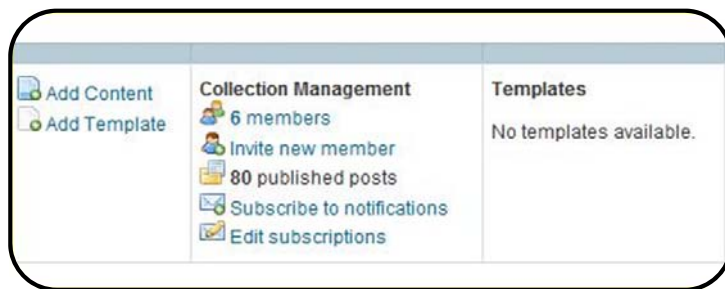
You are now on the Create Metadata page; this is where you need to enter information. There are twelve tabs on the left hand side, all labeled. You will click these tabs to enter information that corresponds with the metadata record.

When you have finished filling in information, making sure everything that is required is completed, you may preview your metadata, or you can immediately save it. You can also publish your information, or if you are not the manager of the collection, you will have to wait for it to be approved by the manager first.

Return to your DASHBOARD and click on My Submissions. Here you will be able to see everything you have submitted. You can click on a metadata record to view, edit, see revisions made, track or clone it.

● OTHER RESOURCES

USGIN Lab

http://lab.usgin.org/

This web site provides information related to service profiles and implementation in development for the US Geoscience information network (USGIN). It is meant to be a site where developers can learn about the standard in use, the objectives of the application profiles being developed, software being used or tested for implementation of services, and details about particular implementations.

Catalog Connector

http://sourceforge.net/projects/catalogconnecto/

Catalog Connector is an open source, Java based CSW client that can search multiple CSW catalogs and CSW implementation versions at once.

GLOSSARY

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.

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.

Content Model: Application to access metadata.

Dataset: an identifiable collection of data. 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.

Dublin Core metadata standard: is a simple yet effective element set for describing a wide range of networked resources.

A Harvester: (harvesting) is operated by a service provider as a means of collecting metadata from repositories.

Metadata record: consists of a set of attributes, or elements, necessary to describe the resource in question. For example, a metadata system common in libraries -- the library catalog -- contains a set of metadata records with elements that describe a book or other library item: author, title, date of creation or publication, subject coverage, and the call number specifying location of the item on the shelf.

Resource: Generic identifiable item of Interest (examples pdf's, data files, physical Rock samples, chemical analyses)

Service Metadata: This is Metadata that describes the operations and information available from a server.

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

Steward: person responsible for updating the metadata.

Web-accessible Folder: is an HTTP accessible directory of files, typically metadata files in XML format in which all files and their time-stamps are visible to a web browser or client. Crawlers are able to parse the file listings and date-time stamps and provide a search interface on these documents.