Information access using the Geoscience Information Network

U. S. GIN Team 2013-04-01T13:24

Contents

[Welcome to the USGIN 3](#_Toc352585992)

[What is USGIN? 3](#_Toc352585993)

[The History of USGIN 3](#_Toc352585994)

[Network components 4](#_Toc352585995)

[Catalog 5](#_Toc352585996)

[Metadata 5](#_Toc352585997)

[Information exchanges 6](#_Toc352585998)

[Content model 6](#_Toc352585999)

[Interchange format 7](#_Toc352586000)

[Community Specifications 7](#_Toc352586001)

[Category: Applications 8](#_Toc352586002)

[Category: Web Service Protocols 9](#_Toc352586003)

[Category: Metadata Standards 9](#_Toc352586004)

[Category: XML interchange formats 10](#_Toc352586005)

[Category: Other Standards 11](#_Toc352586006)

[Data access platforms 11](#_Toc352586007)

[Web applications 11](#_Toc352586008)

[Files 11](#_Toc352586009)

[Web Services 11](#_Toc352586010)

[Using the Network 12](#_Toc352586011)

[Getting Data 12](#_Toc352586012)

[Publishing data 13](#_Toc352586013)

[Tiered data access 13](#_Toc352586014)

[Creating Metadata 14](#_Toc352586015)

[Setting up data service 16](#_Toc352586016)

[Developing applications 17](#_Toc352586017)

[Applications 18](#_Toc352586018)

[Documentation and resources 18](#_Toc352586019)

[Metadata and catalog 18](#_Toc352586020)

[Information Exchanges 19](#_Toc352586021)

[System 19](#_Toc352586022)

[Tutorials 19](#_Toc352586023)

[Glossary 19](#_Toc352586024)

[Terminology 19](#_Toc352586025)

[Frequently Asked Questions (FAQ) 20](#_Toc352586026)

[Who uses USGIN? 20](#_Toc352586027)

[What are USGIN Subdomains? 21](#_Toc352586028)

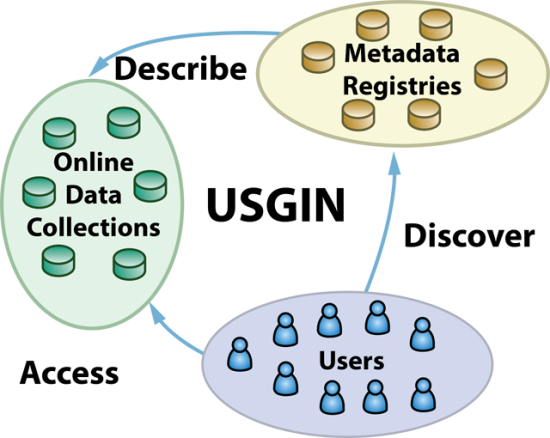
[References 21](#_Toc352586029)

# Welcome to the USGIN

* The United States Geoscience Information Network (USGIN) enables users to efficiently find, access, and share geoscience data, reducing the time and effort spent locating useful information and documenting new data.
* Information is registered and discovered based on standardized catalog services and metadata
* The system accommodates resources in various forms, from unstructured text and images to documented, community web services and interchange formats.

# What is USGIN?

The US Geoscience Information Network (USGIN) can be understood from several perspectives. The simplest view is that the network is the collection of web-accessible resources that are registered in online catalogs conforming to USGIN practice. USGIN can also be considered a network of computer systems built on standard Web architecture, sharing data using conventions for service protocols, interchange formats, and vocabularies to enable geoscience information discovery, access, and usage. USGIN is also a community of users following an approach to geoscience information access that is based on distributed resource providers and a collection of public, open source specifications. From this final perspective, USGIN is a microcosm of the larger Internet, designed to simplify the utilization of geoscience information.



USGIN is a network, and is thus largely intangible. The network is defined by the protocols, interchange formats, and conventions that enable its operation. Thus, the documents that define these protocols, interchange formats and conventions are the artifacts that represent the network. These documents are archived in the USGIN repository (http://repository.usgin.org), indexed through the USGIN Catalog , and accessed using standard Web resource retrieval.

# The History of USGIN

State, federal and private institutions across the United States have accumulated large stores of geoscience data over the last 150 years. These data resources have played a vital role in meeting the energy and mineral resource demands of our society, as well as planning for and mitigating geologic and environmental hazards, and understanding the Earth.

The advent and growth of the World Wide Web as the entry point to obtain information has changed the landscape of public data delivery, requiring Geological Surveys to rethink their mission to make geoscience information available to the public. Representatives of the Association of American State Geologists (AASG) and the U.S. Geological Survey (USGS) met in Denver February 21-22, 2007, to discuss opportunities for making geoscience data more accessible and interoperable across agencies. They recommended that the USGS and State Geological Surveys work together to create a distributed, national “Geological Information Network” (GIN) of digital Earth Science data using common standards and protocols, preserving ownership, credit, and control of data, and building on existing data systems (Allison and Dickinson, 2008).

Funding from the National Science Foundation in 2009-2012 (award number EAR-0753154) supported development of USGIN foundation principals, engagement with State Geological Surveys, and deployment of a catalog system. This work is currently the basis for implementing the DOE-funded National Geothermal Data System (http://geothermaldata.org).

# Network components

The core components of the network are information exchange specifications and the catalog. Information exchange specifications are community agreements on the conventions necessary for the interoperable exchange of some particular information. The exchange specification specifies a scope, the content model for the data items of interest, interchange formats for encoding and transmitting information electronically, and the protocols used to request information. By using an interchange exchange specification a data provider can publish data that will be available to any client that implements the exchange, and a client application can access data from any provider publishing data according to the exchange conventions. The catalog comprises a collection of metadata records that describe resources accessible through the USGIN, and a special information exchange that defines metadata content, how the metadata collection is searched, and how metadata is encoded in search responses.

As a system, USGIN is a loosely coupled system of independent data providers, client applications, and infrastructure. The infrastructure includes 1) tools for registration of new resources, searching the metadata catalog, authentication, and resource validation; 2) registries for vocabularies, agents, specifications, and interchange schema; and 3) documentation and educational resources. Because network operation is based on information exchange specifications that are independent of any particular hardware or software implementation, all of the operational components can evolve as technology evolves. Use of standard protocols enables data access using off the shelf software, both commercial and open-source. Open Geospatial Consortium Web Map Service and Web Feature Service, and OpenDAP NetCDF services are specified in current USGIN information exchanges. Details on the use of these services, and interchange formats for data exchange, are described later in this report.

The distributed nature of the system means that stewardship of resources is determined by resource owner. Participation in the network only requires that a resource provider create metadata that conforms to the USGIN profile, and make the metadata and the described resource available. The network is open; anyone can deploy new nodes and components that implement one or more USGIN specifications. New specifications can be introduced for service protocols, interchange formats, or vocabularies. Keeping resources under the stewardship of the parties responsible for the information promotes system sustainability because the stewards have a direct connection with the quality of the product.

## Catalog

The catalog is first and foremost a collection of metadata records describing resources that are intended to be considered part of the USGIN. The catalog should be thought of as a cloud of records; there may be no individual datastore on the network that contains the entire collection of metadata records. To support this collection (registry) and make it useful, components are necessary to enable data originators to register resources by adding metadata to the catalog, and to enable data consumers to find, evaluate and use the described resources. The fundamental basis for integrating the catalog system is the conventions for metadata content items and how they are used. The use of a standardized encoding scheme for metadata interchange and a standard service protocol for accessing catalog content simplifies interoperability and enables the use of off the shelf software to implement the catalog system. The scope of the catalog includes data products or datasets; these are units of information that have been authored, edited or compiled under some stewardship, with some purpose and procedure, and having a common collection of access processes.

Metadata records constituting the catalog are aggregated by one or more catalog servers that expose the metadata for searching. Current search technology includes exposing metadata records for indexing by commercial search engines, and exposing metadata via the Open Geospatial Consortium Catalog Service for the Web (CSW). Various software applications have been implemented to assist users to search the catalog and present results in a user-friendly manner.

## Metadata

Metadata should be created and submitted to an USGIN catalog for any resource that is meant to be accessible individually. Individual documents require one metadata record per document. These documents might be scans of well logs, scanned reports or publications, or data in a spreadsheet, such as an Excel file. 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; if it is not accessible online, the metadata should provide instructions on how to access the document.

Datasets (data products) are typically considered as individual works (see FRBR, Tillett, 2003), unified by the compilation activity that brings information together into a single data structure, editing and verifying content as necessary. This approach is based on our interest in facilitating data access by users, and the observation that such a user is first interested in a particular kind of data; upon finding a fit-for-purpose dataset, they will next want to know how to get the data. From this perspective a dataset work (data product, resource) will have a single metadata record that may include specification of multiple distributions of the resource. For instance, a borehole temperature dataset may be available with all the records in an Excel spreadsheet, or visualized through a web map service, or individual observations may be accessed through a web feature service (see for example [Montana Thermal Springs](http://repository.stategeothermaldata.org/repository/resource/168566464e3d5f8f3cde3b9fc0049f45/)).

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

The required metadata content is explained in USGIN metadata recommendations (USGIN Standards and Protocols Drafting Team, 2010). These requirements specify the content of the metadata, but not the delivery format. ISO19139 xml is the preferred encoding based on its expanding adoption in the community. USGIN guidelines for implementing the content recommendations in ISO19139 XML (USGIN Standards and Protocols Drafting Team, 2012) are available, as well as a detailed USGIN profile for interoperable metadata using ISO19139.

FGDC xml is widely used and if participants already have workflow in place using this format and can provide the requested metadata content, this can be harvested using an FGDC-to-ISO transformation. FGDC XML should be tested to validate against the official XML schema at http://www.fgdc.gov/schemas/metadata/fgdc-std-001-1998.xsd.

## Information exchanges

An information exchange is a contract that a data provider follows to expose a particular kind of data through a particular service protocol in a particular interchange format. Each exchange specification defines one or more web services. The development of information exchange specifications outlined here is based on operational systems of the Open Geospatial Consortium (OGC, http://www.opengeospatial.org/ogc/process), International Organisation for Standardisation (ISO, http://www.iso.org/iso/home/standards\_development.htm), and the EPA Environmental Information Exchange Network (<http://www.exchangenetwork.net/data-exchange/>), and the National Geothermal Data system (NGDS, http://geothermaldata.org/page/data-interchange-content-models).

### Content model

Content models are abstract specifications that define a feature for information interchange, along with the properties associated with the feature and relationships to other features. The content model in its simplest form is a statement that some particular entity (feature or observation) will have some list of properties. For example a content model for a record describing a book might include: title, author, publisher, publication date, publication place, number of pages, and an ISBN number (a URI).

New content models are developed based on community needs. A network participant may propose a new model for development and form a working group to develop a draft for review and potential adoption. Content models may be developed as Excel workbooks, text descriptions, or UML models. Documents used in the development process are hosted in a version-controlled (VC) repository (https://github.com/usgin-models). Documents defining each model are managed in one repository. The development documents describe and specify the model, but do not implement the model in a computer-processable form. These documents are intended to help human users understand the content model.

Review copies of model specification documents are made available at a web location (typically not the development version-controlled repository) with access control based on the target reviewer community. Availability of a model for review is publicized in a request for comment broadcast to the intended user community, with instructions for commenting, and a time frame during which comments will be accepted. Comments are collected by e-mail and compiled in a comments spreadsheet that is maintained in the content model VC repository. At the end of the review period, the workgroup incorporates comments to produce a final specification document. When the group is satisfied that the model is ready for implementation and utilization, a copy of the specification document is moved to an archival ‘tag’ repository, and locked against further modification.

A URI is assigned to the model at this point; the specification document is considered the normative representation of the model. Ideally the model URI will dereference to the specification document. The URI should include version information (see versioning and URI’s section, below).

### Interchange format

In order to actually use a content model, it must be implemented using an encoding scheme that is computer processable. The eXtensible Markup Language (XML) is typically used to implement the content models, but other schemes (e.g. JSON) may be adopted as technology evolves. The important thing is that the information is encoded consistently in a structured, well documented format. This enables computer programs to parse the interchange documents and extract desired information, and as newer encodings are adopted, conversion between formats can be automated using software.

Test implementation of a content model should be done during the review of the abstract model, because this provides an excellent review of the model. When the specification is finalized (‘tagged’), a final implementation can be developed. The implementation is specified by a separate document. In the case of XML implementations, the normative specification is an XML schema; other encoding schemes will have other normative implementation specifications. The implementation encoding is identified by a URI (a Namespace URI in the case of an XML implementation). This URI should be included in any instance document based on that interchange format implementation such that software consumers of the document content can determine that the document is encoded using a known format.

Each implementation must define a validation process that will determine if an instance document is conformant with the implementation.

## Community Specifications

One of the operating principles of the USGIN is to not reinvent the wheel, i.e. to use and extend existing software whenever possible instead of developing new software. Active, free, open-source software communities offer a path to long term viability of key system components whether or not individual funded projects continue.

USGIN specifications in general define conventions and practices for the use of existing components and standards to simplify interoperability. This page presents a list of the community standards, [profiles](http://usgin.org/glossary#profile), [protocols](http://usgin.org/glossary#protocol), and software employed by USGIN to address the challenges associated with sharing large amounts of diverse data stored in geographically disparate locations. In general, these are not the only solutions available, but indicate the resources that we have found useful.

| Category: [Applications](http://usgin.org/specifications/applications) | | |
| --- | --- | --- |
| [Specification](http://usgin.org/specifications?&order=field_specification_label&sort=asc) | [Description](http://usgin.org/specifications?&order=field_specification_description&sort=asc) | Uses |
| [Amazon Web Services](http://aws.amazon.com/) | Computing infrastructure resources available for rent. [More…](http://lab.usgin.org/applications/amazon-web-services) | Several servers that host USGIN websites or provide USGIN services are provided by Amazon Web Services. |
| [Apache Tomcat](http://tomcat.apache.org/) | Free-and-open-source servlet engine. Many key applications for providing web services can be hosted using Tomcat. [More...](http://lab.usgin.org/applications/apache-tomcat) | Used to host software applications, such as GeoServer and GeoPortal, which provide [Open Geospatial Consortium (OGC)](http://www.opengeospatial.org/)-compliant [web services](http://usgin.org/glossary#web_service) |
| [Django](http://www.djangoproject.com/) | Open-source web application framework written in the Python programming language. [More...](http://lab.usgin.org/applications/django) | Used to develop a number of USGIN web applications, including the USGIN URI redirection engine. [More…](http://lab.usgin.org/groups/using-django-usgin) |
| [Drupal](http://drupal.org/) | Modular, extensible, actively maintained free-and-open-source web site content management framework; capable of running on any PHP-capable web server. [More...](http://lab.usgin.org/applications/drupal) | Websites related to the USGIN Initiative built using the Drupal include:  The [AASG Geothermal Data](http://www.stategeothermaldata.org/) website  The [NDGS Portal](http://geothermaldata.org/), USGIN Lab site  The [USGIN](http://usgin.org/) website. [More…](http://lab.usgin.org/groups/drupal-development) |
| [ESRI ArcGIS](http://esri.com/) | Commercial geographic information system software and web service-compatible server software. Supports both proprietary service protocols ('Geoservices API'), and OGC services. | Commonly used for deploying geospatial data as [web services](http://usgin.org/glossary#web_service). |
| [GeoServer](http://geoserver.org/) | Free-and-open-source web service-compatible server software. [More...](http://lab.usgin.org/applications/geoserver) | Southern Methodist University, a contributor to the [National Geothermal Data System](http://geothermaldata.org/), uses GeoServer to host web services; the landing page is [here](http://geothermal.smu.edu/geoserver/web/). |
| [PostGIS](http://postgis.refractions.net/) | Extension for PostgreSQL relational database to support geographic objects and operations. [More...](http://lab.usgin.org/applications/postgresql-and-postgis) | PostGIS databases are used to persist spatial data for ArcGIS, GeoServer, Django, and many other applications. |
| [PostgreSQL](http://www.postgresql.org/) | Free-and-open-source object-relational database system with over 15 years of active development. [More...](http://lab.usgin.org/applications/postgresql-and-postgis) | PostgreSQL is used by GeoPortal, GeoNetwork, CKAN. |
| [Python](http://www.python.org/) | A programming language. [More...](http://lab.usgin.org/applications/python) | Python is used for USGIN Django sites and converting Excel spreadsheets into ISO metadata (the application responsible for which can be found at <http://github.com/usgin/csvtometadata>). |

| Category: [Web Service Protocols](http://usgin.org/specifications/web-service-standards) | | |
| --- | --- | --- |
| [Specification](http://usgin.org/specifications?&order=field_specification_label&sort=asc) | [Description](http://usgin.org/specifications?&order=field_specification_description&sort=asc) | Uses |
| [ESRI Map Service](http://webhelp.esri.com/arcgisserver/9.3.1/dotNet/index.htm#what_is_a_map_service.htm) | Specification for geographic data web services deployed using ESRI software. Currently under consideration as an OGC standard. [More...](http://lab.usgin.org/standards/esri-map-service) | Many USGIN data services are published using ArcGIS server; because it is simple to deploy the ESRI Map Service along with the OGC services using this server, they are typically deployed in parallel. |
| [OGC Catalog Service for the Web (CSW 2.0.2)](http://www.opengeospatial.org/standards/cat) | Service specification to support publishing and searching collections of geospatial resource descriptions ([metadata](http://usgin.org/content/glossary#metadata)). [More...](http://lab.usgin.org/standards/ogc-catalog-service-web-csw-202) | Catalog services deployed by USGIN |
| [OGC CSW APISO - OGC ISO19115/ISO19119 application profile for CSW 2.0](http://www.opengeospatial.org/standards/cat) | A [profile](http://usgin.org/glossary#profile) of the [OGC](http://www.opengeospatial.org/) [CSW](http://usgin.org/glossary#csw) 2.0.2 standard. Users who wish to deploy an [OGC](http://www.opengeospatial.org/)-compliant [catalog service](http://usgin.org/glossary#csw) for their data should do so in accordance with this profile. [More...](http://lab.usgin.org/standards/ogc-csw-apiso-ogc-iso19115iso19119-application-profile-csw-20) | Users typically deploy [catalog services](http://usgin.org/glossary#csw) to list and manage queries regarding a large number of different web-accessible resources. Compliance with this [profile](http://usgin.org/glossary#profile)of OGC specifications maximizes compatibility with other OGC services and ensures a degree of compatibility with USGIN services. |

| Category: [Metadata Standards](http://usgin.org/specifications/metadata-standards) | | |
| --- | --- | --- |
| [Specification](http://usgin.org/specifications?&order=field_specification_label&sort=asc) | [Description](http://usgin.org/specifications?&order=field_specification_description&sort=asc) | Uses |
| [ISO 19115](http://iso.org/) | Standard for geospatial [metadata](http://usgin.org/content/glossary#metadata); developed by the International Standards Organization (ISO). [More...](http://lab.usgin.org/standards/iso-19115) | Governs metadata records describing datasets and entities. |
| [ISO 19119](http://iso.org/) | Standard for metadata describing geospatial [web services](http://usgin.org/content/glossary#web_service). [More...](http://lab.usgin.org/standards/iso-19119) | Governs metadata records describing web services. |
| [ISO 19139](http://iso.org/) | XML Implementation of ISO 19115 and ISO 19119 standards. [More...](http://lab.usgin.org/standards/iso-19139) | All USGIN-compliant metadata conforms to this standard. |

| Category: XML interchange formats | | |
| --- | --- | --- |
| Specification | Description | Uses |
| [Geography Markup Language (GML)](http://www.opengeospatial.org/standards/gml" \t "_blank) | An [XML](http://usgin.org/content/glossary#xml) grammar for encoding geospatial location and its association with thematic data. Developed by the [Open Geospatial Consortium (OGC)](http://www.opengeospatial.org/) [More...](http://lab.usgin.org/standards/geography-markup-language) | Basis for geospatial location encoding in USGIN [feature](http://usgin.org/glossary#feature) services. |
| [GeoSciML](http://www.cgi-iugs.org/tech_collaboration/geosciml.html" \t "_blank) | GeoSciML is a [GML](#GMLTableEntry) application scheme for application-neutral encoding of geoscience data and related spatial data. [More...](http://lab.usgin.org/standards/geosciml-xml-markup-language) | Geoscience [interchange format](http://usgin.org/glossary#interchange_format) used for OneGeology level 5 services. |
| [GeoSciML-Portrayal](https://www.seegrid.csiro.au/wiki/CGIModel/GeoSciMLPortrayalViewModel) | A simplified implementation of the GeoSciML conceptual model; simple feature encoding for geologic unit outcrops, contacts, faults, boreholes, and data observations. [More...](http://lab.usgin.org/gsmlp) | Use for OneGeology US services. [GeoSciML-Portrayal Cookbook](http://usgin.org/sites/usgin.org/files/GeoSciML-PortrayalCookbook.pdf). |
| [W3C XML](http://www.w3.org/TR/REC-xml/) | A framework for defining application-specific [markup language](http://usgin.org/content/glossary#markup_language)s developed by the World Wide Web Consortium. [More...](http://lab.usgin.org/standards/w3c-extensible-markup-language-xml) | XML is the basis for the specialized [markup languages](http://usgin.org/glossary#markup_language), [interchange formats](http://usgin.org/glossary#interchange_format), and [schemas](http://usgin.org/glossary#schema) used by USGIN; these include GML, GeoSciML, and GeoSciML-portrayal. For more information about XML, see the [USGIN XML Tutorial](http://usgin.org/content/xml-tutorial/) |

| Category: [Other Standards](http://usgin.org/specifications/data-standards) | | |
| --- | --- | --- |
| [Specification](http://usgin.org/specifications?&order=field_specification_label&sort=asc) | [Description](http://usgin.org/specifications?&order=field_specification_description&sort=asc) | Uses |
| [W3C Simple Knowledge Organization System (SKOS)](http://www.w3.org/TR/2009/REC-skos-reference-20090818/) | A standard way to represent knowledge organization systems using the Resource Description Framework (RDF). [More...](http://lab.usgin.org/standards/w3c-skos) | Encoding of vocabularies for GeoSciML documents by the CGI Geoscience Terminology Workgroup. |
| [IETF URI](http://tools.ietf.org/html/rfc3986) | Internet Engineering Task Force (IETF) specification for construction of Universal Resource Identifier strings. [More...](http://lab.usgin.org/standards/ietf-rfc-3986-universal-resource-identifier) | USGIN URIs conform to a USGIN- [profile](http://usgin.org/glossary#profile) of the IETF URI standard. |

## Data access platforms

Utilization of the internal content in datasets may be enabled in several ways. An online web application might be provided that allows filtering, browsing, and even analytical processing of data online. In many cases datasets can be downloaded as files, and the internal content is accessed offline using desktop software. Web-based access to individual features, observations, or granules is enabled in the USGIN by data services.

### Web applications

A web application is a unit of software functionality that is accessed from a web server, and is executed in a controlled environment (for security) by the client’s web browser. The software (php, javascript, python, java) that is executed to run the application is downloaded from the web server when the application is activated. The application may communicate with one or more servers using open protocols, but typically uses a tightly coupled application-specific communication scheme only known to the web application and the servers it is designed to interact with. The application may offer access to one or more resources in a variety of ways, but some sort of form-based querying and browser based visualization (maps, tables, graphs…) is generally the case. Such applications may also offer file downloading to acquire representations of resources in a useful format for the user in their environment.

### Files

The simplest and most common access to resources is provided by simple web links that result in a file download. The content of the file must be in a format that can be interpreted and used by some software in the user’s host environment.

### Web Services

A web service is a unit of functionality with operations that can be invoked using requests sent using the WWW. A web service specification defines an interface, including operations available, the messaging protocol required to invoke operations, and the structure and content of messages used by the service to respond to requests. HTTP is almost certainly the most widely used web service. It defines a simple set of operations (GET, PUT, DELETE, POST) that enable the World Wide Web. Technically, file-based access to resources via HTTP is a web service, but for the purposes of USGIN, data access through web services denotes capabilities that enable filtering or processing of the data beyond those available from HTTP.

# Using the Network

## Getting Data

The basic work flow to access data on the network is like accessing any information on the web. First a search is conducted to locate the information of interest, then that information must be acquired in a form that is useful to the user. Many textual, file-based resources registered in the network will be discoverable using standard search engine technology (Google, Yahoo, Bing, etc…) and web browser-based file access using URLs. The USGIN metadata system is designed to expose such items for indexing by commercial search engines, so that USGIN resources can be discovered using a Google search, e.g. "[Sustainable Geothermal Power](https://www.google.com/search?&q=%22Sustainable+Geothermal+Power%22)".

Data in databases or other data-oriented file formats is not indexed for effective discovery using commercial search engine technology. These information resources must be described in text, with additional information required to understand data formats or web services used to access the data. This is the function of metadata records. A USGIN catalog search application can take advantage of the standard metadata that describes USGIN resources to add value, such as enabling spatial search and directly linking users to view or download specific data that they need from large datasets.

File-based datasets are acquired by familiar download processes. If the content of the file uses one of the standard data exchange schema, it will be relatively simple for a user to extract the information they need. Data in unstructured or ad hoc-structured file formats will typically require additional effort by the user to understand how to extract the information they need in a form they can use. Data that are accessible through standard USGIN web services might be immediately accessed from the user's application environment (e.g. ArcGIS, Excel).

The goal of the USGIN discovery system is to integrate data discovery and access into standard web architecture. Software widgets that search the catalog are not difficult to implement because standard web technology is used, and should be relatively portable across applications. Search results encoded in standard structured metadata allows software to bring data directly into the application environment without having to open web browsers, go to web sites, search for data, download the data, reformat, import, etc. In Web jargon, the discovery system is a hypermedia application that can be integrated into any Web-aware system.

To demonstrate some of the search capabilities and provide an entry point for discovering USGIN resources, use the [USGIN Catalog](http://catalog.usgin.org/search). The [AASG Geothermal Data Catalog](http://catalog.stategeothermaldata.org/), is another catalog application that utilizes the USGIN metadata system, demonstrating a domain-specific catalog instance focused on geothermal data. Catalog search interfaces can also be integrated into other software environments like ArcGIS. An example is the [USGIN CSW extension for ArcGIS](http://lab.usgin.org/applications/doc/csw-client-esri-arcgis-desktop-and-arcexplorer), which searches the USGIN catalog, and can load selected map services from a results list to an ArcMap project. Another example is <http://data.geothermaldatasystem.org/>, which will search the catalog for feature services, and download selected data as comma-delimited text that will open directly in most spreadsheet software.

## Publishing data

Providing data to the US Geoscience Information Network can be as simple as making some files accessible online, or as complex as setting up a full [GeoSciML 3.0](http://www.geosciml.org/geosciml/3.0/documentation/html/index.html) OGC Web Feature Service. In any case, one of the most important parts of the workflow is creating informative and complete metadata to document the resource.

* Creating metadata
* Publishing files [link to page about repositories, file management, naming etc.]
* [Setting up data services](#_Setting_up_data)

For more information, USGIN data providers and data developers are encouraged to peruse the .

### Tiered data access

The USGIN is a unified data access system based on the registration of resources in a shared catalog system using standardized metadata. The system has a tiered data access scheme accommodating file-based, non structured, and standards-based structured data delivered using standardized web services and interchange formats. A data resource becomes part of the system when standard USGIN metadata is created, validated, and made discoverable through the catalog system, and the data resource is accessible via procedures specified in the metadata. Much of the information that is or will be registered is unstructured data. Other resources, such as drill cores, may not be available in electronic format.

The tiered data delivery scheme allows flexibility to accommodate unmanaged legacy data in whatever form it is available, as well as common or high value structured data using documented information exchanges. An informal community governance scheme is used adopt new information exchanges, and the network infrastructure provides the USGIN repository where the specifications for each exchange are available to all.

* **Tier 1: Unstructured** — represents file based resources, unstructured data in text, image, etc.; requires a human to extract data for analysis.
* **Tier 2: Structured, but not standardized** — represent data structured in formats that are not conformant with a registered USGIN content model. Data in this tier would need to be transformed in case by case fashion by the data consumer.
* **Tier 3: Structured, standardized** — data published based on a USGIN information exchange.

A large part of the available resources are scanned images of legacy reports, maps, and other figures that are registered with metadata and made available as Tier 1 resources. Tier 2 allows registration of existing structured datasets in whatever form they are currently available. This is not a preferred data acquisition approach, but is expedient and useful for unique datasets that have only a single instance, or when the cost of transforming to a standard interchange scheme is not justified.

Tier 3 data acquisition is the preferred scheme, but because of the additional effort required to edit and review datasets to get them into the standard interchange format, it will be necessary to prioritize effort.

### Creating Metadata

Metadata for the USGIN catalog might come from a variety of sources. In many cases, there is some kind of existing metadata that can be transformed to import into the USGIN catalog. For individual new resources, it is generally easiest to create metadata using a form interface (e.g. [INSPIRE metadata editor](http://inspire-geoportal.ec.europa.eu/editor/)). The use of standardized metadata and encoding should also improve software-assisted automatic metadata generation as part of normal data processing workflows (e.g. [GeoChron](http://www.geochron.org/reductionsoftware.php), which makes metadata acquisition mostly transparent for very specific laboratory workflows).

Existing metadata compilations include lists of files in a text document or spreadsheet, various databases constructed by organizations to manage their library holdings, and formal metadata conformant in varying degrees to Federal Geothermal Data Committee (FGDC) or rarely ISO standards, constructed according to locally varying interpretation and practice. In some cases, the metadata collected is not sufficient to conform to the USGIN recommendations, and additional content might need to be added.

The challenge to promote better metadata acquisition is to minimize the manual data entry required, making metadata acquisition as simple as possible. Approaches include user-friendly forms, spreadsheet editing that is familiar to most computer users, transformation processes from existing database metadata, and automated metadata extraction. Metadata requirements can be relaxed for many content elements that are mandatory in the interchange schema to allow ‘missing’ as a value for some required content.

Form-based tools can background user log-in information to auto-populate some of the metadata, as well as providing pick lists and autocomplete functions in the data entry fields. The date and timestamp of submission can be recorded, saving the data provider from having to create this data for each submission manually.

Use of spreadsheet for metadata compilation and editing is popular, but does not support one-to-many relationships that are common in metadata content. Spreadsheets allow users to do extract/trans­form/load processing from existing metadata tabled or spreadsheet using familiar cut, paste, search/replace, and fill-down operations supported by spreadsheet software. Spreadsheet metadata-compilation-table columns can be mapped to the USGIN ISO metadata profile, and metadata entered in each row can be automatically converted to an XML record for import into the catalog.

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

#### Location information

In order to enable basic geographic search using a map interface, each resource metadata record must have a latitude-longitude bounding box delineating the geographic area that is the subject of the resource. Metadata creation form interfaces commonly allow the user to draw a rectangular box in a map view. With care this can produce accurate location metadata, but is time consuming, typically requiring 3-5 minutes per metadata record. If this is deemed too much effort, locations can be specified using place-name keywords. In some cases, if there is a good correspondence between a named location (mountain range, valley, known geothermal resource area) and the subject area for a resource; this gazetteer approach can yield good results, and in many cases the named location can be correlated with a geographic bounding box (geocoded) to enable map-based geographic search.

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

#### Automatically Generated Metadata

Some metadata, such as the electronic transfer protocol used to retrieve an online resource (examples: FTP, HTTPS), and the methods required (HTTP Post, Get) can be populate by default when metadata describes a resource that is uploaded to a repository. Structural metadata such as the MIME type, can be inferred during a file upload process as well.

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

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

#### Manual Metadata Entry

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

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

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

### Setting up data service

Review existing [information exchanges](#_Information_exchanges) to determine if there is one already defined for the information you have

publish metadata for service that specifies the expected life time for the service and how notification of service disruption or termination will be made.

#### Exchange Exists

#### No existing exchange

##### [Tier 2](#_Tiered_data_access) publication

Data can be published in whatever form it is in; structured data in non-standard formats

##### Define new information exchange

Members of the USGIN community propose specifications for data sharing exchanges. Proposed exchanges must be documented and reviewed. After review, documentation and any artifacts necessary for data exchange (e.g. XML schema for interchange formats) are made available through the USGIN document repository and schema registry. Conformance with the specifications allows users to share data. More information about USGIN specifications can be found on the USGIN Specifications page.

Setting up a USGIN information exchange:

1. Define content model. Recommend scoping the model based on the intended use case, and on example datasets that the interested parties want to share.
2. Get review of content model from community of expected users. Revise content model as necessary.
3. Select interchange service protocol. This should be based on the availability of server and client softare in the community of users. Some example possibilities include OGC WFS, WMS, WCS, OpenDAP/NetCDF/THREDDS, Microsoft OData, ESRI Geoservices API.
4. Implement the content model using an encoding scheme compatible with the chosen service protocol
5. Define validation rules for instance documents
6. Document the content model, interchange format, service protocol, and any special conventions or profile. Specify how data access links to data exposed using this exchange will be described using the metadata fields in USGIN profile ISO19139 CI\_OnlineResource elements. Use existing identifiers where possible to identify service and MIME types.
7. Deploy an example service and test with client software. Iterate 1-5
8. Have documentation reviewed by target users and technical experts; respond to comments, updating 1-7 as necessary.
9. Register a datasets distributed using the exchange in a USGIN catalog, and deposit specification documents in USGIN repository.

## Developing applications

Software applications that take advantage of the standards based approach to data documentation, discover, and delivery are the actual manifestation of the USGIN. Anyone can develop applications that utilize USGIN services using various commercial and [free-and-open-source](http://usgin.org/glossary#open-source) software (FOSS) platforms.

The [Lab.usgin.org](http://lab.usgin.org) web site is a good place to learn more about the standards in use, and software projects that provide a foundation for USGIN applications. In particular the [Tools](http://lab.usgin.org/tools) page provides a listing of web apps and documents generated by the project to help organize and document USGIN-related development. Various community discussion groups are also active that can provide ideas and assistance. [developer tutorials will be linked from here]

Open-source applications are being created for viewing and exploring data provided by Web Services, motivated by the increasing amount of data accessible by Web Services, particularly when services are [OGC-Compliant](http://www.opengeospatial.org/standards/), making for a system of interoperable data. To join a community of developers and to see how applications are being developed and used to implement NGDS, visit the Tools, Community and Group pages of [USGIN Lab](http://lab.usgin.org/):

* [USGIN Lab tools](http://lab.usgin.org/tools)
* [USGIN Lab Community Homepage](http://lab.usgin.org/usgin-community-homepage)
* [USGIN Lab OG](http://lab.usgin.org/og)

USGIN software development projects, many of which support NGDS functions, are open source, and accessible at GitHub in the [USGIN Github repository](https://github.com/usgin).

## Applications

[USGIN Document repository](http://repository.usgin.org/) (by Wolfgang Grunberg and Ryan Clark). Register document resources, and upload to repository or link to document. Metadata is exported to web-accessible directory for harvest into USGIN catalog. Development notes [here](http://lab.usgin.org/groups/drupal-development/creating-document-repository-drupal).

[USGIN Catalog](http://catalog.usgin.org/geoportal/catalog/main/home.page) is currently implemented using [ESRI Geoportal](http://geoportal.sourceforge.net/) to provide a search interface and CSW service. Metadata is stored in a MySQL database as ISO19139 XML documents. Search is enabled by Lucene indexing against an abstract metadata model that includes CSW core and ISO CSW profile queryable elements.

[GeoSciML WFS client](http://lab.usgin.org/implementations/arcmap-geosciml-wfs-client) proof of concept application (by Averill Cate)

[Catalog client for ArcMap](http://lab.usgin.org/applications/doc/csw-client-application) add-in (by Genhan Chen)

[Django application for URI redirection](https://github.com/usgin/uriredirect/blob/master/readme.md) (by Ryan Clark). Has management console to build rewrite rules using regular expressions. This is used as resolver for http://resources.usgin.org/uri-gin USGIN URI's, but can be easily configured for other URI schemes. Code base has been overhauled as of 04/2012 to make it more portable.

# Documentation and resources

The governance of the network has been initiated by the State Geological Survey in the US, under the umbrella of the Association of American State Geologists and the US Geological survey, with the objective of improving access to and utilization of their information resources.

Introduction of specifications that duplicate capabilities of specifications already in use is discouraged, except when these provide for utilization of newer technology. Specification documents, schema, and model artifacts are listed here. Software development projects are all publicly accessible on GitHub at <https://github.com/usgin>. The current process for adopting specifications is quite informal because the user community is small.

## Metadata and catalog

[Metadata Recommendations for Geoscience Resources](http://repository.usgin.org/uri_gin/usgin/dlio/335): This document provides guidance on the metadata content required to meet the use requirements for USGIN metadata. The content model is not specific to any particular metadata standard or implementation.

[Use of ISO metadata specifications to describe geoscience information resources](http://repository.usgin.org/uri_gin/usgin/dlio/337): This document describes recommended practices for using ISO19139 xml encoding of ISO 19115 and ISO 19119 metadata to describe a broad spectrum of geoscience resources.

[USGIN metadata best practices](http://repository.usgin.org/uri_gin/usgin/dlio/499) Guidelines for implementation of USGIN metadata content recommendations in ISO 19139 XML metadata.

## Information Exchanges

[Simple Feature Content Model Guidelines](http://repository.usgin.org/uri_gin/usgin/dlio/340): guidance for creating an Excel workbook that defines a content model for a flat file data interchange format, to be implemented in XML as a GML simple feature

[State geothermal data content models](http://geothermaldata.org/page/ngds-content-models): this page is a directory of the information exchange content models in use by the [National Geothermal Data System](http://geothermaldata.org/)

## System

[URI policies:](http://repository.usgin.org/uri_gin/usgin/dlio/331) This document presents a proposal for minting of dereferenceable http URIs for use in interoperable web services.

## Tutorials

The United States Geoscience Information Network (USGIN) is a large-scale initiative that mobilizes a number of challenging concepts. In addition, some USGIN specifications can be technical and demanding. This page provides a list of tutorials to describe those concepts and procedures that are most likely to prove challenging.

Though all of the tutorials below reflect USGIN specifications, some are more applicable to the [National Geothermal Data System](http://usgin.org/www.geothermaldata.org), which implements USGIN specifications.

These tutorials are available in PDF form; the PDF can be found on the first page of each tutorial.

* [NGDS Content Model Tutorial](http://usgin.org/content/ngds-content-model-tutorial)
* [NGDS Feature Search and Map Client](http://usgin.org/content/ngds-feature-search-and-map-client)
* [NGDS Naming Conventions](http://usgin.org/content/ngds-naming-conventions)
* [Rasters and Vectors: A Comparison](http://usgin.org/content/rasters-and-vectors-comparison)
* [USGIN CSW Client](http://usgin.org/content/usgin-csw-client)
* [USGIN Metadata Tutorial](http://usgin.org/content/usgin-metadata-tutorial)
* [USGIN URI Tutorial](http://usgin.org/content/usgin-uri-tutorial)
* [XML Tutorial](http://usgin.org/content/xml-tutorial)

In addition to the above, the following tutorials are available in PDF form:

* [AASG Geothermal Data Repository Tutorial.pdf](http://usgin.org/sites/usgin.org/files/sgd-repository_tutorial.pdf)
* [Data Delivery Cycle Tutorial.pdf](http://usgin.org/sites/usgin.org/files/Data_Delivery_Cycle.pdf)
* [GeoSciML-Portrayal Cookbook.pdf](http://usgin.org/sites/usgin.org/files/GeoSciML-PortrayalCookbook.pdf)

# Glossary

## Terminology

web service: a software application that offers a collection of operations that may be invoked via requests sent using http protocol on the world wide web. Requests are sent to a web location that will be referred to as the service end point. A service instance is identified by the URL of its service end point.

Web map service: a web service that provides georeferenced map images. A request will typically identify a ‘layer’ and a map extent. A layer is defined by a particular data set with a particular legend (portrayal scheme) applied to symbolize that data. Services may offer other request parameters to specify different map projection, image format (tiff, png, jpg…), pixel-dimensions of the image, or a custom-defined portrayal scheme. A given map service instance offers at least one layer, but may offer as many layers as the provider chooses.

Web feature service: a web service that provides data describing particular identifiable features (e.g. bridges, buildings, roads, geologic outcrops, faults), encoded in a format that can be transmitted electronically. A common example is the OGC Web Feature Service, which provides XML-encoded information about geographically-located features. The service provider defines the feature types, content model for the feature type, and the encoding scheme for the information provided. The content model defines the information elements (a.k.a. fields, attributes, properties…) associated with a feature. A feature type has exactly one associated content model, but the content associated with a feature type may be encoded in different ways (e.g. GML, JSON, CSV, KML, ESRI shapefile…). A given feature service instance offers at least one feature type with a default encoding format, but may offer multiple feature types and encoding formats.

conformant: an artifact (model, implementation, instance document) implements all normative provisions of a specification exactly. The artifact may implement other features that are not normative in the specification (http://pubs.opengroup.org/architecture/togaf9-doc/arch/Figures/48\_conformance.png

compliant: the system provides support for some of a given standard

instance document: a document containing information of interest encoded according to some interchange format specification.

# References

Allison, M. L., and Dickinson, T.L., 2008, A workshop on the role of state geological surveys and U.S. Geological Survey in a geological information system for the nation: Tucson, Arizona Geological Survey Open File Report 2008-01, 22 p. (accessible at [http://repository.azgs.az.gov/uri\_gin/azgs/dlio/122 2013-02-01](http://repository.azgs.az.gov/uri_gin/azgs/dlio/122%202013-02-01)).

USGIN Standards and Protocols Drafting Team, 2010, Metadata Recommendations for Geoscience Resources: U.S. Geoscience Information Network Best Practices Document, Doc ID gin2010-11, v. 1.0.3, available at <http://lab.usgin.org/profiles/doc/metadata-content-recommendations> (accessed 2010-11-18).

USGIN Standards and Protocols Drafting Team, 2012, Guidelines for implementation of USGIN metadata content recommendations in ISO 19139 XML metadata: U.S. Geoscience Information Network Best Practices Document, available at http://repository.usgin.org/uri\_gin/usgin/dlio/499.

Tillett, Barbara, 2003, What is FRBR? A Conceptual Model for the Bibliographic Universe: Washington, D. C., Library of Congress, Cataloging Distribution Service, 2003. (accessed at <http://www.loc.gov/cds/downloads/FRBR.PDF> 2013-03-31)