Information access using the Geoscience Information Network

U. S. GIN Team 2013-07-15T10:08

Contents

[Abstract 1](#_Toc361642029)

[What is USGIN? 1](#_Toc361642030)

[The History of USGIN 2](#_Toc361642031)

[Network components 3](#_Toc361642032)

[Catalog 4](#_Toc361642033)

[Metadata 4](#_Toc361642034)

[Current guidelines 5](#_Toc361642035)

[Information exchanges 5](#_Toc361642036)

[Content model 5](#_Toc361642037)

[Interchange format 6](#_Toc361642038)

[Community Specifications 7](#_Toc361642039)

[Data Access 8](#_Toc361642040)

[Tiered data access 8](#_Toc361642041)

[Data access approaches 9](#_Toc361642042)

[File-based approach 9](#_Toc361642043)

[Web applications 9](#_Toc361642044)

[Web Services 10](#_Toc361642045)

[Using the Network 10](#_Toc361642046)

[Getting Data 10](#_Toc361642047)

[Publishing data 11](#_Toc361642048)

[Creating Metadata 11](#_Toc361642049)

[Setting up a data service 13](#_Toc361642050)

[Developing applications 16](#_Toc361642051)

[Relation to other cyberinfrastructure 16](#_Toc361642052)

[Governance 17](#_Toc361642053)

[Glossary 17](#_Toc361642054)

[References 18](#_Toc361642055)

# Abstract

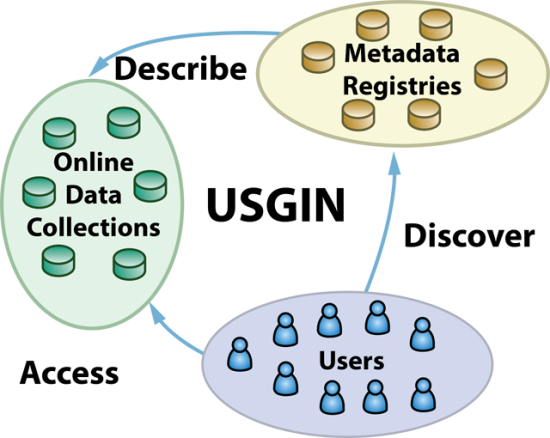
The U.S. Geoscience Information Network (USGIN) is a partnership of the Association of American State Geologists (AASG) and the U.S. Geological Survey (USGS), who formally agreed in 2007 to develop a national geoscience information framework that is distributed, interoperable, uses open source standards and common protocols, respects and acknowledges data ownership, fosters communities of practice to grow, and develops new Web services and clients.

Over the last 2 years, the US Geoscience Information Network (USGIN) vision has materialized with the deployment of the DOE National Geothermal Data System (NGDS). Although much remains to be done, key components are in place and operational, including: 1) Metadata and file repositories; 2) catalog services using standardized metadata; 3) information exchange schemes for a variety of important data types; 4) client applications; and 5) tutorial and explanatory documentation. As of February, 2013 181 Data services are now being hosted by 12 state geological surveys, with 20 feature types implemented. The catalog system is based on the USGIN profile of ISO19115/19139 and the OpenGeospatial Consortium (OGC) Catalog Service for the Web (CSW 2.0.2). Data services are mostly OGC Web Map Service and Web Feature Services. Principal client applications for data consumption are ESRI ArcGIS and Microsoft Excel. In the spectrum of related systems, USGIN occupies a niche characterized by focus on application-neutral delivery of Earth Science data using existing standards and protocols. An open-data mind set is motivated because we are a network of Geologic Surveys largely tasked with the stewardship of non-proprietary geologic data and their dissemination and utilization for the benefit of society. Like many government open-data initiatives, the emphasis on data documentation and accessibility has highlighted the gap in development of demonstration client applications to attract community interest and engagement.

# What is 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 and integrating useful information and documenting new data.
* Information is both registered (by providers) and discovered (by users) 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.

The US Geoscience Information Network (USGIN) can be understood from several perspectives. The simplest view is that the network is a 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.



The USGIN is a 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 tangible expression of the network is the community of users that utilize it, and the resources that are available using its conventions. The result is a large scale flow of information that allows a user to acquire data by searching catalogs that describe resources available from numerous providers.

# The History of USGIN

State, federal, academic, and private institutions and individuals 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. Geological Surveys have mandated missions to not only collect such data, but also to archive and disseminate them. Scientists spend a significant percent of their time finding and transforming data into formats to be integrated for analysis and interpretation. If we can reduce the effort spend on discovery and transformation, we will see a transformative impact on productivity and innovation.

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 envision new ways to make geoscience information more readily available and usable to the end-user and 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 and to the broader geoscience community and other stakeholders and clients. They recommended that the USGS and State Geological Surveys work together to create a distributed, national “Geoscience 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). The GIN agreement was presented to a national workshop the next month as a service of the geological surveys to help facilitate community convergence on a vision for geoinformatics (e.g., cyberinfrastructure for the geosciences). It was widely applauded as an enabling element.

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 US 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 particular units of information. The exchange specification defines a scope, the content model (descriptive components) for the data items of interest, interchange formats for encoding and transmitting information electronically, and the protocols used to request information. These exchange specifications enable a data provider to publish data that will be available to any client that implements the exchange; conversely, a client application can access data from any provider publishing data according to these conventions. This allows data in a variety of formats, organization, and structure to be integrated without having to manually transform it.

The catalog comprises a collection of metadata records that describe resources accessible through the network (in this case, USGIN), and exchange instructions that define metadata content, how the metadata collection is searched, and how metadata are encoded in search responses.

USGIN is a loosely coupled collection of independent data providers and client applications coordinated by a standard infrastructure. The infrastructure includes

* tools for registration of new resources, searching metadata catalogs, authentication, and resource validation;
* registries for vocabularies, agents, specifications, and interchange schema;
* documentation and educational resources.

Because network operation is based on information exchange specifications that are independent of any particular hardware or software, 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 (the data provider). Participation in the network only requires that a provider create metadata that conforms to USGIN exchange specifications, 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, without requiring approval. 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. Metadata records in the catalog describe data products or datasets: 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. 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 (at any node in the system), and to enable data consumers (client applications) to find, evaluate and use the described resources. The fundamental basis for integrating the catalog system is adoption of 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.

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, e.g. http://search.usgin.org, assist users in searching the catalog and present results in a user-friendly manner.

## Metadata

Metadata should be created and added to the catalog for any resource that is meant to be accessible individually. Individual documents require one metadata record per document. These documents might be scanned reports, maps, or publications, multimedia files (e.g., a video or audio recording), or data compiled in a spreadsheet, such as an Excel file. Some document types may consist of a bundle of files, e.g. an ESRI shape file. In general these should be bundled into a single file container 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 Functional Requirements for Bibliographic Records (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 data about a particular subject; upon finding a fit-for-purpose dataset, they will next want to know how to get the data. From this perspective a dataset or data product 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 (WMS), or individual observations may be accessed through a Web Feature Service (WFS) (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 noting 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, date 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 [metadata content model](http://repository.usgin.org/uri_gin/usgin/dlio/335) defines the bounds of information that is expected to be included in metadata records, which in turn determines the granularity of searching that may be done with the metadata catalog.

### Current guidelines

The required metadata content is explained in [USGIN metadata recommendations](http://repository.usgin.org/uri_gin/usgin/dlio/335) (USGIN Standards and Protocols Drafting Team, 2010). These requirements specify the content of the metadata, but not the delivery format. ISO19139 XML is the encoding format adopted for use in the USGIN, and guidelines for implementing the content recommendations in this XML format (USGIN Standards and Protocols Drafting Team, 2012) are [available from the USGIN web site](http://usgin.org/specifications/metadata-and-catalog) (<http://usgin.org/specifications/metadata-and-catalog>).

XML encoded metadata following the FGDC Content Standard for Digital Geospatial Metadata (CSDGM, http://www.fgdc.gov/metadata/csdgm/) 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 [CSDGM-to-USGIN-ISO transformation](https://github.com/usgin/usgin-geoportal/blob/master/WEB-INF/classes/gpt/metadata/fgdc/csdgm2iso19115_usgin.xslt) . 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 set of conventions that a data provider can follow to expose a particular kind of data to be interoperable with other applications using the same conventions. Each information exchange specification defines one or more Web services. The process for information exchange specification development outlined here is based on operational procedures in use by 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), the National Information Exchange Model (NIEM, https://www.niem.gov/technical/model-package-description/Pages/iepd-lifecycle.aspx), the EPA Environmental Information Exchange Network ([http://www.exchangenet­work.net/data-exchange/](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 specifications that define a feature, the properties associated with the feature and relationships to other features. The content model in its simplest form is a statement that a particular feature (or observation) has a set list of properties, some of which are mandatory. 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). Content models might provide a lengthy list of properties but do not require that all of them be provided for any individual record.

New content models are developed based on community needs. Network participants can develop their own content models for proprietary or shared use. The value of a content model will vary depending on how effectively it either fills a gap in the USGIN inventory of models, or supports an alternative service that extends existing discovery and access capabilities for USGIN resources.

A network participant may propose a new model and form a working group to develop a draft for review; the process is defined by "Defining a new USGIN Information Exchange" (draft version at [DefineNewInformationExchange.docx](https://github.com/usgin/usginspecs/blob/master/DefineNewInformationExchange.docx?raw=true)). Content models may be developed in a variety of formats, such as Excel workbooks, text descriptions, or UML models. Documents in the development and review process are hosted in a version-controlled repository (https://github.com/usgin-models). 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, and 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 development 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 made public in a network repository at http://schemas.usgin.org. 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 can be transmitted electronically and processed by computer programs to extract desired information. The eXtensible Markup Language (XML) is currently used to implement 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. As newer encodings are adopted, conversion between formats can be automated using software.

Test implementation provides an excellent review of a content model and should be done during the review period. When the content specification is finalized (‘tagged’), the interchange format can be implemented. 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 interchange format.

## Community Specifications

One of the operating principles of the USGIN is to not "reinvent the wheel". Therefore the project uses and extends 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 projects continue. USGIN specifications describe conventions and practices for the use of existing components and standards to enable interoperability. These specifications can be implemented in a variety of ways. Table 1 presents a list of the resources 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 as a guide for new users.

Table 1. Applications useful for deploying USGIN services.

| **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. [CKAN is used by data.gov] |
| [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>). |

# Data Access

The USGIN is a unified data access system based on the registration of resources in a shared catalog system using standardized metadata. A data resource becomes part of the system when standard USGIN metadata are created, validated, and made discoverable through the catalog system, and the data resource is accessible via procedures specified in the metadata.

## Tiered data access

The system has a tiered data access scheme accommodating file-based, non-structured, and standards-based structured data. Much of the information that is or will be registered is expected to be unstructured data. Some resources, such as drill cores, might not be accessible in electronic format, although metadata for them can be. The tiered data delivery scheme allows flexibility to accommodate legacy (i.e. existing analog records such as print documents and maps) 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 to adopt new information exchanges, and the network infrastructure provides the USGIN system 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** — represents 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. A critical requirement for making Tier 2 data reusable is documentation of the data structure, entities, and attributes in the dataset. Tier 3 data access is the preferred scheme, but the additional effort required to edit and review datasets to get them into the standard interchange format will not be justified in every case.

## Data access approaches

Information resources may be accessed on the Web through a variety of approaches, which can be categorized into file-based, web-application, and web-service approaches. In many cases datasets can be downloaded as files, and the internal content is accessed offline using desktop software. A Web application might be provided that allows filtering, browsing, and even analytical processing of data online. Web services provide granular access to individual features, observations, or granules within larger datasets, designed for machine-to-machine processing.

### File-based approach

The simplest and most common access to resources is provided by simple Web links that result in a file download. The contained information can be accessed by users who have software that can recognize and open files in the format it is delivered. They can utilize the information if they can understand the encoding, language, and data structure, but the system provides no support for this understanding, and little or no automation is possible. This is the standard model for files accessible on the web, supported by HTTP servers and desktop web browser software.

### Web applications

A Web application is a computer program that is accessed from a Web server, and is executed in a controlled environment (for security), typically a Web browser, in the user's computer. The software (e.g., php, javascript, python, java) that is executed to run the application is downloaded from the Web server when the application is activated. The operating 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. Examples include the USGS 'the National Map' (<http://viewer.nationalmap.gov/viewer/>), the Ocean Data Viewer (<http://data.unep-wcmc.org/>), the USGS Water Information System Web Interface (<http://waterdata.usgs.gov/az/nwis/rt>), and the Nevada Bureau of Mines and Geology Geothermal Web Map (<http://gisweb.unr.edu/geothermal/>). These applications provide useful functionality, but do not lend to interoperability or resource reuse, because the application functionality is typically tightly coupled to a particular data source. In such a case, application function cannot easily be applied to other data sources, and the data cannot be accessed directly by other applications.

### Web Services

A Web service is computer program that can be instructed to perform various functions (operations) using requests sent via the World Wide Web (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; file-based access to resources via HTTP is a Web service. The service defines a simple set of operations (GET, PUT, DELETE, POST) that enable the Web. For the purposes of USGIN, data access through Web services denotes capabilities that enable filtering or processing of data, designed for use by computer programs to make data access seamless and mostly invisible to the human operator.

# 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 for 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)".

File-based datasets are acquired by familiar download processes. If the content of the file uses one of the standard data exchange schema (Tier 3), a user can extract the information they need with relative ease. Data in unstructured (Tier 1) or ad hoc-structured file formats (Tier 2) will typically require additional effort by the user to understand how to extract the information they need in a form they can use.

One of the rationales for establishing USGIN is that data in databases or other data-oriented file formats are not indexed for effective discovery using commercial search engine technology. These information resources must be described in text, with additional metadata information required to understand data formats or Web services used to access the data. 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.

The goal of the USGIN discovery system is to integrate data discovery and access into standard Web architecture, such that data accessible through standard USGIN Web services might be seamlessly integrated into the user's application environment (e.g. ArcGIS, Excel). 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 allow 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, this discovery system is a hypermedia application that can be integrated into any Web-aware system.

Use the interface for the [USGIN Catalog](http://search.usgin.org) to explore some of the search capabilities for discovering USGIN resources. The [AASG Geothermal Data Catalog](http://search.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 ArcMap](http://lab.usgin.org/applications/doc/csw-client-esri-arcgis-desktop-and-arcexplorer), which searches the USGIN catalog from within the ESRI ArcMap application, and can load selected map services from a results list to the ArcMap project. Another example is <http://data.geothermaldatasystem.org/>, which will search the USGIN catalog for feature services, and download selected data as comma-delimited text that will open directly in most spreadsheet software.

## Publishing data

Providing data through 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
* Setting up data services

### Creating Metadata

There are many possible workflows to create informative and standardized metadata that accurately describes a resource. 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 for individual resources can be created or updated using a form interface (e.g. http://repository.usgin.org) for metadata generation. Form-based tools can acquire 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.

Metadata can also be compiled in spreadsheet templates (e.g. [http://schemas.usgin.org/­models/­#Metadata](http://schemas.usgin.org/models/#Metadata)) and exported to USGIN ISO XML. The use of spreadsheets for metadata compilation and editing is popular, but does not support one-to-many relationships that are common in metadata content. Examples of such relationships include multiple authors, multiple links to related resources, and multiple distribution options (e.g. file download, order hard copy, web map service). Spreadsheets allow users to do extract/trans­form/load processing from existing metadata using familiar cut, paste, search/replace, and fill-down operations supported by spreadsheet software, but their flexibility also allows users to modify columns or not follow data type conventions that can break automated processing of the content. 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.

In many cases, existing metadata in some form can be transformed to import into the USGIN catalog. Existing metadata compilations might consist of 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 Geospatial Data Committee (FGDC) or to ISO standards. In some cases, the available metadata content is not sufficient to conform to the USGIN recommendations, and additional content must be added. 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.

#### Automatically Generated Metadata

Some metadata information, 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 Web-accessible repository. Structural metadata such as the MIME type, can be inferred during a file upload process as well. 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 a content management system.

Some file types may already contain 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 metadata terms (Dublin Core Metadata Initiative, 2012-06-14) 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/).

#### 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 to delineate the extent of the data content. 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. In many cases named locations 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.

### Setting up a data service

To make data structured in records with a consistent set of attributes accessible via USGIN, the process consists of determining the interchange format that will be used, doing any transformation work necessary to get the data into that format, loading data so it can be accessed by the software hosting the data web service, deploying the service, and registering the dataset metadata in the USGIN Catalog. In any case, the dataset will need to be described in a USGIN conformant metadata record; see the Creating Metadata section for guidance.

To determine if there is an existing information exchange appropriate for the data, review the schemes registered at <http://schemas.usgin.org/models/>. Look at the documentation for the content model to see if your dataset includes the required minimum fields, and to determine that essential information can be accounted for in the scheme. Proceed to 'Exchange Exists' if that is the case; otherwise proceed to 'No Exchange Exists'.

#### Exchange Exists

If an information exchange exists, the content model should be studied carefully to understand each field in the model. The documentation may specify conventions for the units used, how null values are encoded, maximum string lengths, syntax to use for encoding information or other details that may not be obvious. Once the content is well understood, the task is to figure out a workflow to take the dataset in its existing form and transform it to the interchange format. In database systems this process is commonly known as 'extract-transform-load' (ETL), and there are as many different approaches to the problem as there are kinds of datasets. Some common approaches are mentioned here as examples.

Current USGIN content models are 'flat file' formats (all fields have simple data types that are numbers, dates, or strings), and can be described using spreadsheet workbooks in which a spreadsheet table can be used as a template for the data-delivery format. See the template workbooks associated with content models at http://schemas.usgin.org for examples. One of the simplest ETL approaches is simply cutting data from an existing spreadsheet column and pasting it in the corresponding interchange-format-template column. In some cases spreadsheet calculation functions can be used to convert measurement units to those required by the interchange, to concatenate string content from multiple fields into a free text field, or to reformat date strings. Because some data items (e.g. source information) may be the same for every record in the dataset, the fill-down function offered by most spreadsheet software is handy for copying values to every record.

If the existing data are in a relational database, a common approach is to set up a template table in the database with the interchange format fields. If and interchange format template is available in a spreadsheet, that can be imported to generate the template table in the database. This may require exporting the spreadsheet as a text file (e.g. CSV—comma-delimited text) to import as a table in the database. The template table in the database can then be used as the target of an SQL INSERT query to create the records for the interchange document. The query can be constructed to join multiple tables and do a wide variety of sophisticated calculation from fields in the source data to the fields in the interchange format.

If the dataset consists of features that are geolocated as points, the ETL operations can all be done using spreadsheets or databases that are not geographically enabled—the X, Y, and Z (if present) coordinates of the points can be carried as fields in the data. These coordinates will need to be transformed into a 'geometry' type field when the interchange dataset is loaded into the database or file that will support the online Web service that publishes the data. Since most of the interchanges specify that locations should be reported using geographic coordinates (latitude, longitude) in the WGS 1984 (EPSG 4326) spatial reference system (SRS), it may be necessary to transform source coordinates into this reference system. These operations will require some geographic information system functionality.

If features in the interchange format are lines or polygons, with geometry that requires a more complex data structure than 3 numbers, the geometry of the features will need to be managed using a spatially enabled database (e.g. PostGIS, Oracle Spatial, or an ESRI Geodatabase). The thematic content can still be managed separately, but a database key must be available that allows the thematic content to be merged with the geometric features when the data are loaded for the Web service deployment.

Once the data are transformed into a spatially-enabled table or file conforming to the interchange format, that table or file will need to be connected to the Web service software to deploy the service. See Web service deployment, below.

#### No existing exchange

If an information exchange is not available for your dataset, there are two options: 1) publish it as a file in its existing format (Tier 2); or 2) define a new interchange format and use that.

##### Tier 2 publication

The simplest solution is to publish the dataset in its existing format. The data are structured, and can be processed by computer software, but the field names, data types, measurement units, etc. are all likely to be unique to that dataset. Any software that uses the dataset will have to be customized. People wishing to use the data will have to be provided with sufficient information to understand the fields in the data, and will have to spend time figuring out the data structure. If the data are associated with geospatial locations, the dataset can be published as a Web service (see Web service deployment, below). In any case the dataset can be packaged in a file and made accessible on the Web. The use of a text-based file format that can be imported into a variety of spreadsheet or database software is recommended. Application-specific binary formats are the most likely to become unusable in a relatively short time and should be avoided. Comma or tab-delimited text files are common and can be imported by many kinds of software. No matter what publication choices are made, metadata must be created that describes the dataset and its structure in enough detail that future users can understand how to use it. USGIN minimum metadata requirements are trivial. Without them, the data are essentially useless. Our goal is to encourage the most robust metadata possible, to increase ease of discovery, access, and usability.

##### Define new information exchange

The decision to define a new information exchange should be based on the likelihood that others will want to publish similar datasets in the future. The governance of the network has been initiated by the State Geological Surveys in the US, under the umbrella of the US Geoscience Information Network (USIN) Steering Committee, jointly chaired by the Association of American State Geologists and the US Geological Survey, with the initial objective of improving access to and utilization of their information resources. The proposal and adoption process at this point is very informal. The intention is to keep the barrier to entry low, based on the philosophy that it's better to have the interchange formats that people wish to use presented in a public forum, with any discussion documented for future reference, and to motivate people to document the formats they are using for future reference and to promote re-use.

Members of the USGIN community propose specifications for data sharing exchanges. Exchange documents are developed and reviewed using a publicly accessible repository on GitHub ([https://github.­com/­usgin-models](https://github.com/usgin-models)). Each exchange has a separate repository associated with the usgin-models gitHub profile. Our workflow guidelines require that a proposed model must have an identified steward, and a working group of at least three participants with relevant domain knowledge and understanding of the interchange technology. There is no formal process for defining workgroup membership; normally the challenge is finding a sufficient number of qualified individuals to provide meaningful reviews and comment. The exchange steward is responsible to assemble the workgroup and assure sufficient expertise in the group to generate a sound content model and implementation. The exchange steward requests creation of a new model repository at the usgin-models gitHub from the organization members, and identifies workgroup members who will have authorization to make changes in documents in the repository (committers). Any community member can create a repository branch to propose changes using standard GitHub procedures, and request consideration for merging back into the developing model.

After review and approval by the workgroup, a call goes out to the USGIN technical review e-mail list for comments from the community; an open review period of 4 weeks is normal. Comments are recorded using the GitHub issue tracker. Comments from the community must be addressed by the workgroup using written responses in the issue tracker. When issues are resolved to the satisfaction of a majority stakeholders (workgroup and engaged community), the exchange specification is adopted by the workgroup. The final step is editorial review by the web master for http://schemas.usgin.org to verify that all components of the exchange documentation are present and sufficient to provide the necessary guidance and artifacts to implement the exchange.

When a specification is adopted, all associated documents are copied to a 'tag' branch in the gitHub repository by the workgroup, and are not changed after they are 'tagged'. The specification documents are also copied to the exchange repository at <http://schemas.usgin.org>, which is a Web site set up to provide public access to exchange specifications and any related xml schema documents or other artifacts required for the deployment and operation of the information exchange. A more detailed discussion of establishing a new USGIN information exchange is provided in a separate document (see <http://usgin.org/specifications/information-exchange>).

#### Deploying Web services

Datasets that are to be deployed as geospatial Web services must be in a spatially enabled table (e.g. PostGIS, ArcGIS geodatabase) or file (e.g. shapefile, GML file). Various software packages are available for deploying the Web services, including commercial products like ESRI ArcGIS and Snowflake, and free, open-source software like Deegree, GeoServer, and Minnesota MapServer. Although the details of the service deployment vary between the software systems, in all cases the server must be linked with the data that are to be served, a configuration document needs to be defined for the open-source solutions, and for Web map services, the map legend symbolization scheme must be implemented. If OGC services are being used, the capabilities document for the service may need to be edited to provide a complete description of the service. If the dataset is published through a Web service, it is important that the service description specifies the expected life time for the service and how notification of service disruption or termination will be made.

## Developing applications

Software applications that take advantage of the USGIN standards-based approach to data documentation, discover, and delivery are the ultimate purpose of the USGIN. Anyone can develop applications that utilize USGIN services using various commercial and free-and-open-source software (FOSS) platforms. The [Lab.usgin.org](http://lab.usgin.org) Web site provides resources to learn more about the standards in use, and software projects that provide a foundation for USGIN applications. The [Tools](http://lab.usgin.org/tools) page provides a listing of Web apps and documents generated to help organize and document USGIN-related development. Various community discussion groups are also active that can provide ideas and assistance. USGIN software development projects are open source and accessible at GitHub in the [USGIN Github repository](https://github.com/usgin). Studying the code for the various projects there is probably one of the best ways to get started building applications to use the network.

# Relation to other cyberinfrastructure

The USGIN has been emerging in a milieu of cyberinfrastructure development in all realms of human activity. The network framework is based on concepts that have made the World Wide Web wildly successful, including the ideas of loose coupling between components, distributed management of content, openness of the development process, and adaptability to new ideas. USGIN occupies a niche characterized by focus on application-neutral delivery of Earth Science information using existing standards and protocols. A significant amount development effort is devoted to developing and documenting interchange formats in the context of service protocols defined by the Open Geospatial Consortium (OGC). The processes and objectives for developing the interchange formats are similar in many respects to the National Information Exchange Model (NIEM, http://niem.org) and the Environmental Protection Agency Exchange Network (<http://www.epa.gov/exchangenetwork/>). USGIN information exchanges have also defined service protocols used to transport information encoded in adopted interchange formats, which contrasts with these other exchange networks.

The USGIN approach to definition of information exchanges has been to define simple packages of data with a relatively narrow scope, with the use of URIs to link related packages on the web. The most widely used interchange is for metadata for resource discovery, evaluation and access, and this is really the unifying factor for the network. Tier 1 and Tier 2 data (See Tiered data access) delivery follow the standard patterns that have emerged for data publication and sharing on the web. Tier 3 delivery requires data integration on the server side, along with more sophisticated validation of the interchange documents and service protocols with respect to the data interchange specification. The additional work must be justified for the data provider based on value to their user community.

# Governance

Tier 3 data delivery requires an infrastructure for developing and maintaining information exchange specifications. The proposal and adoption process for specifications at this point is very informal. The intention is to keep the barrier to entry low, based on the philosophy that it's better to have the information exchanges that people wish to use presented in a public forum, with any discussion documented for future reference, and to motivate people to document the protocols, content models, and interchange formats they are using for future reference and to promote re-use. Introduction of specifications that duplicate capabilities of specifications already in use is discouraged, except when these provide for utilization of newer technology. Software development projects are all publicly accessible on GitHub at <https://github.com/usgin>.

An open-data mind set is motivated because USGIN is a network of Geologic Surveys largely tasked with the stewardship of non-proprietary geologic data and their dissemination and utilization for the benefit of society. The governance of the network has been initiated by the State Geological Surveys in the US, under the umbrella of the US Geoscience Information Network (USIN) Steering Committee, jointly chaired by the Association of American State Geologists and the US Geological Survey, with the initial objective of improving access to and utilization of their information resources.

# Glossary

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

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

Federal Geographic Data Committee (FGDC): organization that sets standards for geospatial data produced by federal agencies in the U.S. and consequently has become a more widespread standard.

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

Web feature service (WFS): 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.

Web map service (WMS): 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 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.

# 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)).

Dublin Core Metadata Initiative, 2012-06-14, DCMI Metdata Terms: accessed at <http://dublincore.org/documents/dcmi-terms/> (2013-07-01).

National Information Standards Organization, 2013-02-20, The Dublin core Metadata Element Set: ANSI/NISO Z39-85-2012, ISSN:1041-5635.

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)

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, 2010-11, Use of ISO metadata specifications to describe geoscience information resources, Doc ID gin2010-009, available at <http://repository.usgin.org/uri_gin/usgin/dlio/337>.

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>.