The U.S. Geoscience Information Network

Authors:

Stephen M. Richard, authors TBD (USGIN Standards and Protocols Drafting Team?)

Creation date:

2013-02-17

Last revision date:

2013-03-31 2:43 PM

Document Status:

draft

Publisher:

Arizona Geological Survey

Description:

This document is a summary and description of the US GeoScience Information Network (USGIN). Target audience is geoscience community, publication in EOS?

Identifier:

tbd

Revision History

| **Version** | **Date** | **Comments** | **By** |
| --- | --- | --- | --- |
| draft | 2013-02-18 | Initial document; still needs abstract, work on Governance, How To be a part, and current resources sections (last 3). Needs figures, examples… | Stephen Richard |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Contents

[Introduction 3](#_Toc352394553)

[History 3](#_Toc352394554)

[The value of the network 4](#_Toc352394555)

[Network framework 4](#_Toc352394556)

[Catalog 5](#_Toc352394557)

[Metadata content 5](#_Toc352394558)

[Recommended metadata content 7](#_Toc352394559)

[Metadata interchange format 9](#_Toc352394560)

[Links 10](#_Toc352394561)

[Catalog system 12](#_Toc352394562)

[Data access 12](#_Toc352394563)

[Access platforms 12](#_Toc352394564)

[Web applications 12](#_Toc352394565)

[Files 13](#_Toc352394566)

[Web Services 13](#_Toc352394567)

[Resources and their representations 13](#_Toc352394568)

[Data access tiers 14](#_Toc352394569)

[Content models 15](#_Toc352394570)

[Interchange format 15](#_Toc352394571)

[Service protocols 16](#_Toc352394572)

[Information exchanges 16](#_Toc352394573)

[System repositories and governance 17](#_Toc352394574)

[How to be part of USGIN 17](#_Toc352394575)

[Current Online Resources 17](#_Toc352394576)

[The Future 18](#_Toc352394577)

[References 18](#_Toc352394578)

# Introduction

The United States Geoscience Information Network (USGIN) is a community of practice formed around a commitment to making data available through a system of documented service protocols and interchange formats, with resources registered in a standardized catalog system accessible through web interfaces. Agreement on these standard procedures and formats reduces the time and effort required to locate existing resources, document new resources, and utilize data. The basic tenets of the network are that it 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. Our goal is to make geoscience information more accessible not just to geoscientists but to the broader community of environmental and life scientists. The Network is a partnership of the Association of American State Geologists (AASG) and the U.S. Geological Survey (USGS); USGIN development has been funded by the National Science Foundation (NSF), the Department of Energy through a connection with the National Geothermal Data System, and the USGS through a partnership with the ScienceBase project.

The USGIN is implemented based on public specifications for interfaces between network nodes, a distributed catalog of information resources, and documented practices for information exchange. It is a network layered on top of the World Wide Web, with a loosely-coupled design intended to allow the independent evolution of different system components. The architecture is intended to leverage Web architecture to form an information ecosystem simplifying data discovery and analysis so that system users can focus on research, collection of new data, and data synthesis for new knowledge, and data stewards can focus effort on maintenance, preservation, and data delivery. Standardized data access protocols and interchange formats allow software developers to focus on functional capabilities and usability, rather than data access and transformation.

The Geoscience Information Network can be understood from several perspectives. The simplest view is that USGIN is the collection of resources that are registered in catalogs conforming to USGIN practice. USGIN can also be considered a network of computer systems built on standard Web architecture, identified by the collection of service protocols, interchange formats, and vocabularies those computers use to interact with each other in order to implement end-user functionality for geoscience information discovery, access, and usage. USGIN may also be viewed as 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 essentially a microcosm of the larger Internet designed to simplify the utilization of geoscience information.

## History

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.

## The value of the network

The intention of the US Geoscience Information Network is to benefit the geological surveys and society by reducing the cost of online data publication and access provision, and through easier (lower cost) access to public-domain geoscience data in support of environmental, resource-development, and hazard mitigation planning and decision-making. The operating supposition is that by sharing resources for system development and maintenance, and standardizing data discovery and access mechanisms, the cost of data access and maintenance will be reduced (see Shapiro, 2000). A study by the German Institute for Standardization concluded that the economic benefits of standardization range between 0.2 and 0.9% of the gross national product (DIN, 2000; Blind et al., 2011). These studies focused on standardization in a wide variety of business domains, and we suggest that they apply in the informatics domain as well. Although anecdotal, consider how the music industry landscape has changed with standardized file formats and metadata schemes for recordings, or the seamless connection of most printers to computers using standard interfaces and interchange formats. Standardized access to rich data resources will create business and collaborative opportunities. Development and use of shared protocols and interchange formats for data publication will create a market for user applications, facilitating geoscience data discovery and utilization for the benefit of society.

# Network framework

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. 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 requirements, how the metadata collection is searched, and how metadata is encoded in search responses.

USGIN operates as 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, and authentication; 2) registries for vocabularies, agents, specifications, and interchange schema; and 3) validation tools, 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. The use of standard protocols likewise 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, and to enable data consumers to find, evaluate and use the 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 content

The USGIN content model for metadata enables network users to efficiently find, evaluate, understand, and trust USGIN resources. The scope encompasses all geoscience resources, related to the solid earth, oceans, atmosphere, and the 'critical zone' at the boundaries between these environments where most biological activity takes place. Detailed, domain- and experimental procedure-specific documentation of resources is a deeper problem and cannot be fully address by a generalized, cross-domain metadata scheme; in many cases such metadata is better considered part of the content model for individual data items.

Discovery scenarios for the USGIN must also consider the current state of search technology. Commercial web search engines are extremely good at indexing text-based, hyperlinked, natural language documents, and open-source software such as Lucene or Solr provides excellent text-indexing capabilities in existing metadata catalog implementation. However, standard text- and link-based web indexing by search engines does not work with non-linguistic information resources. Non-linguistic resources must be associated with textual metadata that can be indexed for search and discovery as a proxy for the resource. Although some part of this metadata can be automatically generated as part of processing workflows or by parsing non-linguistic content, scientific data requires human engagement to produce data description, provenance, quality, and other documentation necessary to enable reuse of the data in the future or by users in other communities. This is the focus of the USGIN content recommendations.

To address metadata requirements, the USGIN design team formulated a list of user scenarios that the metadata must support and based metadata content requirements on those scenarios (USGIN, 2011a). These requirements are summarized in the Tables below. Table 1 summarizes what are considered the minimum content elements required for useful metadata. Table 2 summarizes the recommended content for USGIN metadata, which extends the minimum requirements to define metadata content for a functional, distributed catalog system that supports discovery, evaluation, and access to resources. The level of description is targeted to datasets or data products.

Table . Minimum metadata content recommendations. The follow list includes the minimum required content for basic resource description, discovery, and access; not all user scenarios can be implemented with this content, but it is proposed as a good cost-benefit balance. Explanation of fonts used: required (content should be provided), conditional (content required under certain conditions), and optional metadata content. Cardinality: the number of values that can be specified for an element ( in gray); 'not nilable' indicates the metadata record may be rejected if the element does not contain useful content.

| **Content element** | **Cardinality** | **Scope notes** |
| --- | --- | --- |
| **Title** | 1 **entry (not nilable)** | Succinct (preferably <250 characters) name of the resource. The title should be unique within the scope of the containing catalog. |
| **Access Instructions** | 1 **entry (not nilable)** | Description of how to access the resource. Include URL if the resource is accessible online. May be free text, but if the resource is available online, a machine-actionable link should be provided (e.g. atom:link, gmd:CI\_OnlineResource). |
| Description | 1 **entry** | Text description to inform the reader about the resource's content, context, and provenance. |
| Geographic Extent - Horizontal | 1 **entry** | Location specified with either a point or minimum bounding rectangle. Values given in decimal degrees using the [WGS 84](http://en.wikipedia.org/wiki/World_Geodetic_System) datum. Required if resource has location Some resources may not be usefully described by an extent, in which case the keyword 'non-geographic' should be assigned. Note that a point location implies that the scope of the resource is actually restricted to a horizontal extent with an area similar to or smaller than the point location resolution. |
| **Originators** | 1 **to many** | Authors, editors, or corporate authors/curators of the resource. The party responsible for the intellectual content of the resource. |
| **Publication Date** | **1 entry** | Publication, origination, or update date (not temporal extent) for the resource. Use a "year" or [ISO 8601 date and time](http://en.wikipedia.org/wiki/ISO_8601" \l "Combined_date_and_time_representations) format. Alternative date formatting must be machine readable and consistent across all datasets. If no publication date is known, estimate the publication date range, enter the oldest year as the publication date, and include the estimated date range in the Description field. |
| **Distribution Contact Party** | 1 **entry** | The party (name of organization, person, or position/role) to contact about accessing the resource. |
| **Distribution Contact Email** | 1 **entry** | How to contact the party responsible for distribution |
| **Metadata Date** | 1 **entry** | Last metadata update/creation date-time stamp in [ISO 8601 date and time](http://en.wikipedia.org/wiki/ISO_8601" \l "Combined_date_and_time_representations) format. This may be automatically updated on metadata import if a metadata format conversion is necessary. |
| **Metadata Contact Party** | *1* **entry** | The party (name of organization, person, or position/role) to contact with questions about the metadata itself |
| **Metadata Contact Email** | 1 **entry** | How to contact the party responsible for metadata content and accuracy |
| **Metadata Specification** | 1 **entry** | Identifier for metadata specification used to create a metadata record encoding this content. |

### Recommended metadata content

This section extends and elaborates the content requirements with recommended content to produce useful metadata to describe resources, credit the originator of the resource, and inform users how to obtain or access a resource. The resource description should provide sufficient information to assist in discovery of the resource through an online search, and to allow users to evaluate the fitness of the resource for an intended purpose.

Table . Recommended metadata content for USGIN resource description. Includes minimum content fields from Table 1 with additional properties for implementation of a complete data discovery, evaluation, and access. system. The table is divided into sections for metadata content related to resource description, access instructions, and metadata record maintenance. Explanation of fonts used: *Terms in italics are groupings of metadata properties*; required (content should be provided), conditional (content required under certain conditions), and optional metadata content. Cardinality: the number of values that can be specified for an element ( in gray); 'not nilable' indicates the metadata record may be rejected if the element does not contain useful content.

|  |  |  |
| --- | --- | --- |
| Resource description elements | | |
| **Title** | 1 **entry (not nilable)** | Succinct (preferably <250 characters) name of the resource. The title should be unique within the scope of the containing catalog. |
| **Description** | 1 **entry** | Inform the reader about the resource's content as well as its context. |
| **Originators** | 1 **to many** | Authors, editors, or corporate authors/curators of the resource. Ideally, a role property (author, editor, curator, compiler…) would be associated with each originator. |
| **Publication Date** | **1 entry** | Publication, origination, or update date (not temporal extent) for the resource. Use a "year" or [ISO 8601 date and time](http://en.wikipedia.org/wiki/ISO_8601" \l "Combined_date_and_time_representations) format. Alternative date formatting must be machine readable and consistent across all datasets. If no publication date is known, estimate the publication date range, enter the oldest year as the publication date, and include the estimated date range in the Description field. |
| *Geographic Extent - Horizontal* | 1 **entry** | **Location specified by** point or minimum bounding rectangle. **North Bounding Latitude, South Bounding or Point Latitude, East Bounding Longitude, West Bounding or Point Longitude.** Values given in decimal degrees using the [WGS 84](http://en.wikipedia.org/wiki/World_Geodetic_System) datum. Some resources may not be usefully described by an extent; in which case the keyword 'non-geographic' should be assigned. If a particular encoding scheme requires a bounding box, a minimum bounding rectangle will be created if only a point coordinates is given. Note that a point location implies that the scope of the resource is actually restricted to a horizontal extent with an area similar to or smaller than the point location resolution. |
| *Contact - Author or Intellectual Originator* | 0 **to 1 entry** | The primary party responsible for creating the resource.Requires a name (one of **Organization Name**, **Person Name,** or **Position Name), and** contact information (either a Telephone number or an E-mail address)**.** |
| **Bibliographic Citation** | 0 **to 1 entry** | Full bibliographic citation if the resource has been published. |
| **Subject Keywords** | **0 to many** | Thematic, spatial and temporal free-form subject descriptors for the resource. A keyword may be assigned on metadata import if none are present. If possible, submit keywords in separate Thematic, Spatial, and Temporal keyword categories. |
| **Resource Language** | **0 to 1 entry** | Use three letter [ISO 639-2 language code](http://www.loc.gov/standards/iso639-2/langhome.html) (defaults to "eng" for English). |
| **Resource ID** | 0 **to many entries** | Resource identifier(s) following any public or institutional standard. Identified consists of an identifier string and if applicable a **Resource ID Protocol** identifier string that specifies the protocol for the resource ID standard. For example: undefined, ISBN-10, ISBN-13, ISSN, URN, URI, IRI, DOI, HTTP, SSN, etc. Examples: doi:10.1000/182; isbn:0-671-62964-6; issn:1935-6862; azgs:OFR-10-02. Many protocols build the identifier for the protocol into the identifier string. |
| *Geographic Extent – Vertical* | *0* **to 1 entry\*** | **Datum Elevation**, **Datum Type**, **Maximum Elevation, Minimum Elevation**. Values given in meters. Maximum and Minimum Elevations are relative to the reported datum elevation, which will typically be the Earth surface at the location of the resource or sea level. Datum Elevation must be reported relative to mean sea level (MSL) in meters using [EPSG::5714 geodetic parameters](http://www.epsg-registry.org/report.htm?type=selection&entity=urn:ogc:def:crs:EPSG::5714&reportDetail=long&style=urn:uuid:report-style:default-with-urn&style_name=OGP%20Default%20With%20Urn&title=) (WGS 84). Datum type must be a controlled vocabulary (Earth surface, MSL, Kelly bushing, etc.). The maximum is always numerically greater than the minimum elevation. For boreholes with datum at the earth surface, depth below surface is reported as a negative number. **\***Vertical extent may be reported relative to different datum (e.g. sea level, Earth surface) in the same record. Example: core from borehole at depths between 100 and 470 feet (30.3 and 142.4 m), borehole collar at 4787 feet (1450.6 m) above sea level. Vertical extent could be reported in either of the following ways: {0, “MSL”, 1420, 1308} or {1450.6, “Earth surface”, -30.3, -142.4}. |
| *Temporal Extent* | 0 to 1 entry | Temporal range over which the resource was collected or is valid. If the resource pertains to specific named geologic time periods, those terms should be entered as keywords (preferable as part of Temporal Keywords). **Start Date** (**0 to 1 entry**), **End Date** (**0 to 1 entry**; required if start date exists), use [ISO 8601 date and time](http://en.wikipedia.org/wiki/ISO_8601" \l "Combined_date_and_time_representations) format (YYYY-MM-DDTHH:MM:SSZ). |
| **Quality Statement** | 0 **to 1 entry** | Text specification of the quality of the resource. |
| **Lineage Statement** | 0 **to 1 entry** | Text description of the resource's provenance. |
| Resource access instructions | | |
| Access Statement | 1 **entry** | Text instructions for how to access the resource. NOTE—either an Access Statement or a Link to Resource must be provided or the metadata record can be rejected. |
| Link to the resource | 0 **to many** | A URL that enables access to the resource. **URL**, **Link** **Function**, **Representation Format**. URL is minimum content required if a link is included. Optionally, a Link Function term from the ISO19115 OnlineFunctionCode controlled vocabulary specifies what a HTTP GET using the URL will invoke. The link might return an html page, electronic document in some other format, an end point for a service, an online application that requires user interaction, etc. Representation Format is a controlled vocabulary term specifying the format (MIME media types) of a file-based response if applicable. |
| *Distribution Contact* | 1 **entry** | The party to contact about accessing the resource. Requires a name (one of **Organization Name**,Person **Name,** orPosition Name**),** and contact information (either a Telephone number or an E-mail address)**.** |
| **Constraints Statement** | 0 **to 1 entry** | Text that describes the resource's legal and usage constraints. |
| **Distribution Keywords** | **0 to many** | keywords describing the physical form of the resource (core, rock sample, digital file, book, journal article), formatting of resource content (file format, e.g. tiff, xls, MIME type), or physical distribution media (film, floppy disk, online service, hard copy). Table 6 in USGIN ISO metadata profile includes a vocabulary for distribution format for use with the ISO19115 distributionFormat name property. Use of these keywords allows users to search for particular kinds of artifacts. |
| Metadata maintenance information | | |
| **Metadata Date** | 1 **entry** | Last metadata update/creation date-time stamp in [ISO 8601 date and time](http://en.wikipedia.org/wiki/ISO_8601" \l "Combined_date_and_time_representations) format. This may be automatically updated on metadata import if a metadata format conversion is necessary. |
| *Metadata Contact* | 1 **entry** | The party to contact with questions about the metadata itself. Requires a name (one of **Organization Name**, **Person Name,** orPosition Name**), and** contact information (either a Telephone number or an E-mail address). |
| **Metadata Specification** | 1 **entry** | Identifier string for the metadata specification used to create a metadata record encoding this content. Should indicate the base standard and version, as well as any profile that applies to the content or encoding. Ideally the identifier could be dereferenced to obtain information about the applicable specification. Identifiers for metadata encoding specifications to be used in the USGIN system will have to be formally defined and registered for such identifiers to be broadly useful. |
| **Metadata UUID** | 0 **to 1 entry** | A Universally Unique Identifier ([UUID](http://en.wikipedia.org/wiki/Universally_Unique_Identifier)) will be assigned during the metadata import process if one is not provided. Unique identification of each metadata record is required to avoid duplicate entries across multiple metadata catalogs. The UUID format provides unique identification without centralized coordination. |

## Metadata interchange format

In order to deploy an interoperable catalog system the metadata content model must be implemented in a standard encoding for serialization and transmission on the internet. USGIN has adopted a metadata interchange format that is a profile of the ISO 19139 XML implementation of the ISO 19115 metadata standard. This choice was based on the comprehensive scope of ISO 19115, and the wide international adoption of the standard as implemented in XML according to ISO 19139.

A key tenet of the catalog system is the idea that the metadata can be harvested freely between catalog nodes, and the more widely accessible a metadata record is, the more likely the resource it describes is to be used. Use of the USGIN conventions for metadata content and interchange format is the foundation for an interoperable discovery system.

### Links

One of the key objectives of USGIN metadata is to enable automated connection of a user's software application to a web service provided some desired dataset. In the metadata content model this information is abstracted in the 'access instructions' element (Table 1), but in the interchange format, there must be very clear syntax and usage conventions for how link information is encoded to make it useful for automating data access. For links that locate online documents accessible using standard browser and file type resolution technology, the link can be as simple as a single URL element that retrieves a representation of the resource. There are many other kinds of related resources that a resource-access link element may point to, including web services that provide access to a dataset resource, metadata for related or source resources, specifications for standards or extensions to standards.

For data distributions through web services, links in metadata must be accompanied by sufficient description that the linked resource can be accessed and provided to the user automatically, with little or no operator intervention other than selecting the resource in a user interface. The link description in the metadata must be compatible with services using a variety of architectures, including hypermedia-based REST (Fielding [2008]; Amundsen [2011]), component-based remote procedure calls, and object-based URI schemes with basic CRUD operations (Amundsen, 2012-12-14). A content specification for such machine-actionable links is discussed in a separate [USGIN discussion paper](http://lab.usgin.org/groups/metadata-interest-group/actionablelinks) (USGIN, 2012), based on a review of requirements and various solutions to implementing machine actionable links, including the CI\_OnlineResource element in ISO 19115, atom:link (IETF RFC-4287) or the web link element described in IETF RFC-5988. Table 3 summarizes the content elements that need to be associated with links to enable software clients to automate resource access using the link.

Table . Content model for machine-actionable links in USGIN metadata.

| **property** | **scope notes** |
| --- | --- |
| linkage (syn: href, targetURI) | URI that identifies the resource that is the target of the link. This is generally an http URI, which will be dereferenced. The associated attributes provide guidance for client software to determine if it wants to dereference this identifier and what representations is can expect when it does. |
| rel | Semantics of link. Semantics in this context means calculable (see discussion in Coyle, 2010 p. 19). URI from IANA **rel** vocabulary should be included for consistency with IETF RFC-5988. Recommendation is to use the Terms not namespace qualified, following guidance in Atom Specification RFC-4287, section 4.2.7.2. |
| title | Free text to label link in user interfaces. Optional. The content of the "title" attribute is Language-Sensitive. Entities such as "&amp;" and "&lt;" represent their corresponding characters ("&" and "<", respectively), not markup. Link elements MAY have a title attribute. The "title" parameter MUST NOT appear more than once in a given link-value; occurrences after the first MUST be ignored by parsers. |
| type | MIME type of response. Specifies file format and optionally the native software application environment. Intention is that if a type is listed here, it is known to be offered by the host that the href accesses. Note that this is only a hint; for example, it does not override the Content-Type header of a HTTP response obtained by actually following the link. There MUST NOT be more than one type parameter in a link-value; occurrences after the first MUST be ignored by parsers. (default value text/html) |
| protocol | Connection protocol to be used e.g. http, ftp, dns, smb, nfs, smtp, pop. See IETF registry at http://www.rfc-editor.org/rfcxx00.html. Protocol operating at the 'bottom' of the application layer of the OSI network protocol stack. Also allow other identifier schemes: ARK, DOI, EAN13, EISSN, ISBN, ISSN, ISTC, LISSN, LSID, UPC, URN, etc from IETF URI scheme registry; URIs using these other schemes are only dereferenceable if the client knows a priori how to process them. Protocols specific to particular applications that are layered on this base protocol are indicated using serviceType (and version) and the outputScheme. Optional, default is HTTP. |
| serviceType | URI that identifies a service protocol. This specifies protocols for network layer above http/ftp e.g. CSW, WFS, WMS, OpenSearch, OpenDAP, etc. Should be a URI that can dereference to some kind of service specification document. URI for serviceType may indicate a version, but if not, one or more versions that may be requested can be explicitly indicated. Syntax to identify versions is defined by the service specification. Only necessary for remote procedure call type services that overload HTTP requests to invoke other operations. Optional. |
| outputScheme (profile) | profile for content of message retrieve by href URL; URI for xml schema or JSON scheme, other description of data structure and content. Clients look at this to pick the link that will get a representation they can use; unknown outputSchemes can be ignored. This is the information scheme in the layers on top of the MIME type encoding and serviceType; note that the same output scheme might be encoded using different MIME types or accessed through different serviceTypes, so the two are somewhat orthogonal. Optional. |
| Other properties that may be useful (all optional) | |
| altRel | Other domain-specific terms specifying semantics of link, not from IANA registry. (Optional). Multiple values are separated by comma. Rel value string MUST be quoted if it contains a comma (","). |
| altTitle | String that encodes title value in a different character set, and/or contain language information as per [RFC5987]. |
| descriptionURL | detailed text description of what the online resource is/does. Since is not considered good practice to put extensive text in an element attribute, implement by reference with a URL for an HTML description page. |
| media | Indicates intended destination medium or media for style information (see Le Hors et al., 1999, Section 6.13 http://www.w3.org/TR/html401 ). Example values include 'screen', 'tty', 'print', 'braille', 'aural'... Vocabulary appears to be related to type of device or material manifestation (including paper as a device...) that is intended target for resource representation. Default to 'screen', and it is anticipated that other values would be only rarely required. [debate point—is this necessary for links that are intended to operate on the web?] |
| length | Indicates an advisory length of the linked content in octets; it is a hint about the content length of the representation returned when linkage identifier is dereferenced |
| hreflang | describes the language of the resource pointed to by the linkage attribute. When used together with the rel="alternate", it implies a translated version of the entry. Multiple "hreflang" parameters on a single link-value indicate language options that may be indicated by the client. |
| behavior | A comma separated list of properties specifying behavior expected in client when link is actuated. See Table 7 for list of values. |

## Catalog system

The catalog system is a collection of servers that expose metadata records that conform to the USGIN metadata specification. A server might simply offer a collection of metadata records in a web-accessible directory for harvest by other catalogs, or a CSW end point that responds to search requests from other servers. A resource is registered in the network when metadata describing the resource becomes available in a USGIN catalog. The system consists of the metadata collection, server components for creating, maintaining, storing, and indexing metadata records, and components for matching search criteria to metadata records. Client applications provide interfaces that enable users to find, evaluate, and acquire USGIN resources by searching or browsing the metadata catalog.

The USGIN has adopted the OGC Catalog Service for the Web (CSW) version 2.0.2 (Nebert et al., 2007-02-23) as the service protocol for catalog search. The interchange format for metadata content is based on the ISO 19115/ISO 19119 standards, encoded in XML according to ISO19139 using the USGIN ISO metadata profile (USGIN, 2011). Use of standard protocols for searching metadata catalogs enables client applications to be reused by multiple portals, and allows individual search client instances to search multiple catalogs simultaneously.

# Data access

The metadata catalog indexes data products and datasets; access to the internal 'records' within datasets might be enabled in various ways, as described in the next section (Access platforms). Discussion of data access in USGIN continues with a discussion of resources and representations on the WWW, followed by description of the USGIN approach using content models and interchange formats. Service protocols that support both access to data and metadata are discussed in the following section.

## 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 commonly 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.

## Resources and their representations

The term resource is used here to mean an identifiable entity of interest (Berners-Lee et al., 2005-01). For the purposes of the USGIN, resources include physical resources (books, people, rock samples, faults), abstract resources (concepts, theories, hypotheses, imaginary entities), and electronic resources (computer files, software applications). In web architecture, the distinction of a resource and its representation(s) commonly creates confusion about the significance of an identifier. In the case of physical resources, the distinction can be clear: if a rock specimen is the resource, a file containing a tiff image of a photograph of the specimen is a representation. In the electronic world, the distinction can become blurred; definition as a resource (the target of an identifier) or a representation of a resource is context dependent.

Most of the USGIN architecture is concerned with electronic resources that are representations of some physical or abstract entity, for instance a borehole or a permeability measurement. These electronic resources are typically packaged in files, which may be structured or unstructured. Structured electronic resources contain information that is systematically organized to enable machine processing of the content. A simple example is tabular data in a comma-delimited text file. Unstructured electronic resources consist of bit streams that can be converted into images or sound that is meaningful to humans, but generally difficult or impossible to use for machines. Examples include recorded speech, free text, scanned photographs or maps.

Various logical approaches have been developed to build structured representations of the Earth for use in computer systems. Continuously varying properties are well represented by arrays of numbers that specify the value of some physical quantity at a regularly spaced lattice of sample points (a grid). This approach is commonly used for quantities like the magnitude of the Earth's magnetic field, wind velocity, or temperature. To bring order to the chaotic heterogeneity of the Earth, scientists commonly define features—discrete, bounded regions that are characterized by a consistent set of attributes. Geologists define geologic units as a starting point to unraveling the geologic history and structure of a region. Environmental scientists define biomes to analyze ecosystems. Meteorologists define air masses to gain understanding of weather systems. An important example of this feature-based approach to representing the Earth is the Open Geospatial Consortium General Feature Model (Kottman and Reed, 2009-01-15), which is the foundation for feature-based information exchange in USGIN. The key properties of a Feature are that it has a spatial-temporal location and has an identity.

The Observation model provides an alternative approach to data, from the perspective of the individual observation or measurement events that are the basis for quantifying our understanding of the Earth. Observations (Cox, 2010-11-10) represent individual measurements of one or more properties of some real-world phenomena. An observation is an event, in which the value of a property characterizing some aspect of the Earth is determined by means of some procedure. A Feature typically summarizes the results of multiple observations to characterize something like a fault, a geologic unit, a well, a power plant, or a geothermal area. Observations represent the more granular, ‘raw’ data like individual temperature measurements, chemical analyses, or heat flow determinations. Observations may have composite results; for instance an individual well log is considered an observation result from a log run event, and a satellite image can be considered the result of an observation of the light reflected from the Earth surface.

## Data access tiers

USGIN uses a tiered data delivery scheme that allows the flexibility to accommodate unmanaged legacy data in whatever form it is available, as well as high value data in standardized content models and/or interchange formats.

* Tier 1: unstructured, file-based resources, e.g. narrative text, images, tables in documents, etc.; requires human to extract data for analysis.
* Tier 2: structured, but not standardized, data in proprietary formats that are not conformant with a community content model.
* Tier 3: structured and standardized data published using the standardized protocols and interchange formats supported by a USGIN content model (an information exchange).

A large fraction of the available resources consists of 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 that are not in standard content models and interchange formats. This is not a preferred data acquisition approach, but is expedient and useful for unique datasets that have only a single instance. Data in this tier would need to be transformed in some fashion by a data consumer in order to integrate with USGIN-standard datasets. To be useful outside of the community in which it was originally acquired and packaged, Tier 2 data requires metadata that completely describes the data schema and vocabulary.

Tier 3 data acquisition is the preferred scheme, but requires additional effort to edit and review datasets to get them into the standard interchange format. This extra effort will need to have some motivation; one of the challenges of moving towards an open data system is to build this motivation into community practice.

## Content models

The basis for information interchange is a content model that specifies data items that represent entities and their attributes of interest for some community. This conceptual model for the information associated with one or more data items is defined by a specification document. Current USGIN content models, mostly defined in the context of the National Geothermal Data System (NGDS), are specified by Microsoft Excel workbooks. Each content model defines one feature, observation, or grid type designed for some particular information interchange use cases.

Because the content models are designed for information interchange in a service-based environment, they are different from information models designed for data storage and maintenance systems. The operation of the OGC WFS services is such that the content model of the features offered by a service also determines the content that is available for filtering data offered by the service. Thus the content models are typically denormalized to facilitate selecting subsets of a dataset through service requests.

A version of a content model is defined by the collection of fields, cardinality obligations for those fields, and the data types assigned to the fields. A major-minor numbering scheme is used to document content model versions. Major numbers indicate a model based on the same collection of use cases and approach to representation of a feature. Minor versions are incremented when fields are added or removed, obligation rules are changed, or data types are changed.

## Interchange format

As with the metadata system, in order to be used in a web-based information system, content models must be implemented in an interchange format that allows serialization of data for distribution on the internet. A number of international efforts are under way to develop specifications for interchange of geoscience information. These include (GeoSciML) (Richard and CGI Interoperability Working Group, 2007; see also http://geosciml.org), and the OGC observation and measurement model (Cox, 2010-11-10). These models are very flexible and allow representation of a wide range of content, but are thus correspondingly complex and difficult to use. Thus, in the initial phase of USGIN deployment for the National Geothermal Data System and OneGeology US, content models have been defined using relatively simple schema in which property values are specified only by string or numeric-valued elements (no nested or complex data types). The content models are designed to be compatible with the more complex and comprehensive models mentioned above to the degree that is practical.

USGIN is currently implementing interchange formats as GML Simple Features (van den Brink et al., 2011-05-11), compatible with the service protocol in use (OGC WFS). WFS can be consumed by existing clients like ESRI ArcGIS Desktop and Quantum GIS. As clients are developed to utilize richer content, more complex, information-rich interchange formats will become useful. USGIN interchange formats are implemented using the eXtended Markup Language (XML); the syntax and structure of interchange documents is specified by an XML schema.

Multiple implementations of a content model may exist, using different formats, e.g. XML, JSON, CSV…, and an implementation of a particular content model in a particular format may be modified (different field order, field names, data types…). In order to document a service to enable software clients to determine that they are operating on a known document format, any interchange document should include an identifier for the interchange format used by that document. The interchange format identifier must map to a particular version of the interchange feature content model, and documents with the same interchange format ID must validate according to a fixed, documented collection of validation rules (e.g. XML schema, Schematron). USGIN mandates that a namespace URI is defined for each XML schema, and that this URI provides a unique mapping to a particular content model and the implementation version.

# Service protocols

Terminology

The term ‘web service’ will be used here to mean 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.

A map service is 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.

A feature service provides data describing particular identifiable features (e.g. bridges, buildings, roads, geologic outcrops, faults), encoded in a format that can be transmitted electronically. 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. USGIN feature services are currently utilizing the OpenGeospatial Consortium Web Feature Service, which mandates that feature content is encoded using Geography Markup Language (GML), which is an XML application scheme.

# Information exchanges

USGIN operates based on a collection of information exchanges. Each of these is defined by a scope, purpose, service protocol, content model, interchange format, validation processes/conformance tests, and any conventions necessary to enable interoperability for some particular kind of information. The foundation of the network is the metadata exchange based on the OGC CSW and USGIN-profile ISO metadata that enables registering and discovering network resources.

New information exchanges may be proposed by any interested party. The only requirement is that the provisions are clearly documented, and the intended community of users has opportunity to review and comment on the proposed design. Adoption of an exchange specification is market-driven—if the specification is found useful, people will use it. Specification documents must be publicly available through the USGIN repository.

# System repositories and governance

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 made publicly available at <http://usgin.org/specifications>. 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.

Conventions, best practices, specifications, and standards….

# How to be part of USGIN

Network participation is open to any service provider that deploys a USGIN information exchange and registers it in a USGIN catalog, or to any data provider that creates USGIN-profile metadata describing their data holdings, and publishes that metadata to a USGIN catalog.

What is a node (use NGDS definition, when that is settled).

how to deploy and register service

# Current Online Resources

USGIN currently is operating with a CSW catalog, several search applications, a collection of content models, a USGIN repository, and a system website for accessing specifications and educational material, and a community web site catering to developers.

<http://usgin.org>

<http://lab.usgin.org>

<http://repository.usgin.org>

<http://catalog.usgin.org>

Vocabularies

# The Future

Nothing in the USGIN model is unique or untested. The intention of defining the network is to specify a working model for an operational distributed network, not to propose a new way of doing things. This approach to data publication and access supports the ongoing evolution of open data systems, and is predicated on decoupling the stewardship and publication of data on the web from the applications that utilize the data. There is a break with standard software design that couples data storage design with a particular user application scheme; with the hazard of the 'build it and they will come' failure. The long range goal is to start viewing data as a product of the research enterprise, as much as the conclusions drawn from data.

# 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)).

Amundsen, Mike, 2011, Building Hypermedia APIs with HTML5 and Node: Sebastopol, CA, O'Reilly Media, e-book.

Amundsen, Mike, 2012-12-14, Three Common Web Architecture Styles: web page, <http://www.layer7tech.com/blogs/index.php/three-common-web-architecture-styles/>, accessed 2013-02-18.

Berners-Lee, T., Fielding, R., Masinter, L., 2005-01, Uniform Resource Identifier (URI): Generic Syntax: IETF RFC 3986. (accessible at <http://tools.ietf.org/html/rfc3986>)

Cox, Simon, editor, 2010-11-10, Geographic Information: Observations and Measurements OGC Abstract Specification Topic 20, v2.0.0: Open Geospatial Consortium Inc., Document 10-004r3, same as ISO 19156. (accessed at http://portal.opengeospatial.org/files/?artifact\_id=41579 2013-01-24)

De La Beaujardiere, Jeff, editor, 2006-03-15, OpenGIS® Web Map Server Implementation Specification, v. 1.3.0: Open Geospatial Consortium Inc., Document 06-042.

Fielding, Roy T., 2008-10-20, REST APIs must be hypertext driven: web blog and discussion, http://roy.gbiv.com/untangled/2008/rest-apis-must-be-hypertext-driven (accessed 2013-02-08).

Kottman, Cliff, and Reed, Carl, editors, 2009-01-15, The OpenGIS® Abstract Specification Topic 5: Features, v. 5.0: Open Geospatial Consortium Inc., Document 08-126.

Nebert, Douglas, Whiteside, Arliss, and Vretanos, Panagiotis, editors, 2007-02-23, OpenGIS® Catalogue Services Specification v. 2.0.2, Corrigendum 2 Release: Open Geospatial Consortium Document OGC 07-006r1. (accessed at http://portal.opengeospatial.org/files/?artifact\_id=20555, 2013-01-24)