# Data Delivery from NGDS nodes to NGDS users: Metadata, Interchange formats, and XML schema

## **Introduction**

This discussion concerns the process of making information resources available on the internet through the US Geoscience Information Network (GIN). Data delivery as discussed here is from the perspective of network nodes delivering data to network users. The processes discussed here are the same that are being used for the DOE geothermal data system, and are intended to be used by all contributors to the system. The term ‘information resource’ will be used as a blanket concept for any kind of digital resource that is made available via the internet at some URL, including file-based documents, file-based databases, downloadable software applications, web applications that are accessed via a browser, streaming data (video, audio, sensors…).

An important concept necessary to understand this discussion is the distinction of a resource and its representations. In the most general sense, a resource is any identifiable thing of interest. A resource may be physical (a person, rock sample, book, map…) or abstract (a concept, an interpretation of geologic structure, the content of the Declaration of Independence, the melody of a tune). In the case of physical resources, we can interact materially with the resource itself. In the case of abstract resources, we can only interact with a ‘representation’ (manifestation) of the resource accessible to our senses of sight, smell, touch, hearing and taste. On the World Wide Web, we can only obtain representations of resource that can be transmitted electronically, which effectively limits us to sight and sound based representations. As used here, the term ‘information resource’ refers to the kinds of representations we can obtain electronically, generally restricted to mean through the digital encoding system that is the foundation or our computer technology and the internet. The information resources are the ones of interest that are intended to be registered, discovered, and used through the GIN.

There are two major factors that must be considered when deciding how to make a resource accessible through the network. One is the access method, and the other is the usability of the offered representations. Access method options for an information resource include a bundled product, typically a file that can be downloaded, a service endpoint, or through a web application. The bundled product approach needs little additional explanation; it should be familiar to any internet user. The basic model from the point of view of a web browser is click on a link and GET a file; the web browser (or other requesting application) may know how to utilize the file, or may simply ask the user to save the file on their local system. Service endpoints offer an ‘interface’; a contract that specifies a collection of requests that will invoke actions (operations) by the server, and specifies the responses that will be returned by the server as a result of the requests. Typically the client application is using the responses directly to produce whatever result is was designed for (an airplane reservation, displaying a map, locating a restaurant…), and the user need never directly interact with the responses. The web application approach provides a web location that activates an application in the user’s web browser environment, commonly by providing the client with software (php, javascript, python, java) that is executed in a controlled environment (for security) by the client’s browser. This application may communicate with one or more server using a variety of open protocol or, more typically, using a tightly coupled application-specific communication 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.

The usability dimension of the information resources relates to how easy or difficult it is for a client to actually utilize a resource. Usability depends on a variety of factors that include file formatting, information encoding, data structures, and terminology, which are technically referred to as syntax (format, encoding), schema (data structure), and semantics (terminology). Which of these factors are important depends to some degree on the kind of resource. The file formatting and encoding are key aspects for text, image, sound and video recordings that will mostly be view and interpreted by people, but the language (semantics) is also a major factor. For resources containing information intended for machine processing, all three factors come into play. The idea is that if an information resource uses syntax, schema and semantics that existing applications are already capable of interpreting, it will be easier to use than one that is unique. If a unique data schema or terminology is used by an information resource, it is more usable if the schema and terminology are documented with easily accessible and complete metadata.

Syntax has to do with how information is encoded in a symbolic form that can be translated to bits a computer can process. A commonplace analogy is the alphabet and punctuation used to represent language in print. It is possible to use a Cyrillic or Greek alphabet to write English words, but this would be incomprehensible (not usable) to most English speakers, because they learn to write with a Roman alphabet. English language may be written using the Roman alphabet, but if the standard rules of punctuation are not followed, the meaning may be incomprehensible or interpreted erroneously by the reader. These are syntactic considerations. In computer terms, this translates to using common file formats (e.g. shp, dbf, xls, doc, txt, xml, json, pdf, tif, jpg, etc.) and character encoding (ASCII, UTF-8…).

A data schema defines how information is organized to make it interpretable. In language this is known as grammar, but for information processing purposes, consider how a standard spreadsheet table is constructed. The schema defines the attributes (columns) associated with each data instance (row). Any description of a thing follows some schema. For instance, the schema for a ‘person’ might include attributes of age, hair color, eye color, height. A different schema for a person might include as attributes political affiliation, income range, religious affiliation, and ethnicity. The schema also includes provisions for the domain of the attributes, i.e. the valid acceptable values. For instance a person’s age might be specified using terms like infant, preschool, pre-teen, teenage, etc, or using a numeric representation of the number of years since birth. In computer terms, this translates to providing information in a structured format, using data structures that are already in use, or are documented publicly available models (entity-relationship, UML).

Semantic usability is based on understanding the meaning of the entities, attributes, and attribute values in whatever schema they are presented. The language analogy here has to do with understanding the meaning of words, which in natural language involves a word and its context for usage. For instance the semantics of ‘mean’ vary—context is required to understand when the symbol ‘mean’ corresponds to the concept of nastiness, semantics, or average, e.g. “It’s mean to think that I have the means to determine what mean means, but I know it is not the mean of several usages”. In the data context, semantics is implicit in the use of documented names for entities (objects, features) and attributes in a schema, and in the use of controlled vocabularies in the specification of attribute values. A controlled vocabulary is a documented mapping between a collection of symbols and meanings that are generally conveyed to people through text definitions. Although they are getting better, computers are not as good as resolving term ambiguity based on context as are humans; thus for computer processing, concepts are associated with identifiers (URIs) that are unique to the concept, not the common-language terms used to label the concepts.

## **Data delivery options**

Following the above discussion, a network node has three options for providing access to an information resource:

1. Place the information resource at a permanent, publicly accessible web location (repository), and submit metadata to a NGDS-conformant catalog that is linked to the network. The metadata must meet the validation constraints for USGIN metadata (see …)
2. Implement a web service that provides access to the resource at the base data (e.g. record, feature, object) level. The web service can be deployed on any publically accessible server.
3. Set up a web application accessible as a web page through a browser that provides access to the resource.

Any of these access approaches may offer representations with varying degree of usability. The simplest for the data provider is unstructured free text, or scanned images of documents, or audio or video recordings. This meets the most basic level of making information resources available. This simply requires a file that is interpretable by the users software.

The next level is to provide information resources in a structured form, i.e. using some known syntax and organizing and labeling the content (in a data schema) so that it can be parsed and information extracted by software. The utility is greatly increased by included documentation of the information schema, and may be increased more by using a ‘standard’ (documented and widely used) schema.

The most sophisticated and complete information publication uses structured syntax and schema, and includes associations of all words (symbols) used in the resource with explanation of their meaning.

Within this framework, the intention is to set the bar as low as possible for considering an information resource part of the system, while requiring sufficient structure and description to make the resource useful. To this end a resource will be considered part of the network when it is discoverable using an NGDS catalog search, and accessible via the web according to procedures described in the metadata record obtained from the NGDS catalog. This dictates a tiered resource categorization scheme

1. Unstructured resource
2. Structure, but not documented
3. Structured and documented schema
4. Structured, documented schema and semantics

In any case, a resource must be registered in the catalog to become part of the system. This requires production of a metadata record that meets the content requirements outline in USGIN geoscience metadata recommendations (ref ). Currently metadata is required to be made available at level 3—it must be structured using a documented schema. The preferred structure is to use ISO 19139 metadata as recommended in USGIN recommendation for ISO metadata.

Note that consideration should be given to how a resource is located (step 1 above, ‘permanent web location’). It has become apparent in the life of the web that the URL at which a resource is accessed is likely to change over time. To account for this problem and minimize broken links, it has become common practice to provide http URI’s for resources. The URI identifies the resource, and can be ‘dereferenced’ to get a representation of the resource. This is a lookup processes that introduces a level of indirection. The http URI is associated (using standard web architecture) with a server (redirector) that maps the URI to the current URL for a file or a service end point for the resource, and forwards requests for the resource to that endpoint to have it returned to the requesting client. HTTP content negotiation can be implemented as part of this scheme to allow a client to request a representation (html, xml, json, rdf/N3, TIFF…) of the resource that is most appropriate for its requirements.

## **Metadata**

Metadata for the catalog should be created and submitted for any resource that is meant to be accessible individually via the web. The level of granularity for metadata record may require careful consideration of likely user scenarios for metadata consumers.

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

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

The required metadata content is explained in [USGIN metadata recommendations](http://lab.usgin.org/profiles/doc/metadata-content-recommendations). These requirements proscribe the content of the metadata, but not the delivery format. ISO19139 xml is the preferred encoding based on its expanding adoption in the community. Recommendations for metadata encoding using the ISO19139 XML for USGIN metadata are documented in [Use of ISO metadata specifications to describe geoscience information resources](http://repository.usgin.org/uri_gin/usgin/dlio/337). FGDC xml is widely used and if participants already have workflow in place using this format and can provide the requested metadata content, this can be made to work. Software tools exist for producing and editing metadata in both formats. Please confer with USGIN developers about metadata creation to facilitate import of metadata into the USGIN catalog.

URI and URL

In this discussion, the acronym URL (universal resource locator) is used to denote a web location that is the access point for some resource. URI (universal resource identifier) is used to identify a resource. See USGIN URI Policies for a full discussion of the intention of a URI and syntax for constructing URIs recommended by USGIN.