

# NGGDPP Metadata Preparation Guide (01/2013)

*“For inclusion in the National Digital Catalog, metadata shall include certain minimal data describing the sites from which data and samples were collected.”*

The National Digital Catalog does not attempt to precisely describe every nuance of each sample-level data record but seeks to include “minimal data” necessary to distinguish and describe the samples and encourage further exploration of similar items by identifying contacts and, where available, linking to online resources providing additional information about collections and records.

Searches within the National Digital Catalog are thus limited to the information found in this minimal metadata. While the metadata is minimal the importance of each element is significant and it is critical to understand each element and how it will be used in the National Digital Catalog.

## Overview of Metadata Preparation

1. Understand the required and optional metadata accepted by the National Digital Catalog.
2. Map the National Digital Catalog metadata to the sample-level properties within your own existing data holdings.
3. Select a data upload format to use for upload to the National Digital Catalog.
4. Extract your existing holdings into the selected upload format.
5. Upload these extracted files into the National Digital Catalog.

Each of these steps is explained in detail below. Additional information is included at: <http://datapreservation.usgs.gov/catalog.shtml>

## 1 – National Digital Catalog Metadata Elements

The tables in this section list the thirteen (13) metadata elements accepted by the National Digital Catalog. Some metadata elements are mandatory and must be included to adequately describe records. Other metadata elements are optional, and when present, enhance metadata records for National Digital Catalog users.

## REQUIRED METADATA

<sup>1</sup> CollectionID may be viewed in ScienceBase ([www.sciencebase.gov](http://www.sciencebase.gov)) by navigating to NGGDPP Catalog, logging into the Catalog, and selecting appropriate state and physical collection folder: <https://www.sciencebase.gov/catalog/folder/4f4e4760e4b07f02db47dfb4>. The URL for the Collection folder is the CollectionID. For example, IL field notes collection: <https://www.sciencebase.gov/catalog/folder/4f4e49d8e4b07f02db5df141>

<sup>2</sup> An extended list of data types may be found in following document: <http://datapreservation.usgs.gov/docs/NGGDPPMetadataProfile.pdf>

<sup>3</sup> Number of element occurrences provided in record metadata. 1-N – indicates multiple entries are acceptable.

Required Element Name	Definition	# <sup>3</sup>
CollectionID <sup>1</sup>	A unique collection ID assigned by the National Digital Catalog to identify distinct collections. This field is required but may be left blank and assigned during the file loading process within the National Digital Catalog.	1
Title	The human-readable title for the individual record that will be used in any listing or search result. Title should be short for display purposes but contain enough information to distinguish from other records. Examples: Sample: Geologic Sample 160580 Sample: ID: NMDF52900064 TITLE: ISAACS BROS. LEAD-SILVER MINE; Sample: Core Research Center, Cutting DD18216; Sample: Core sample from well: KNIK ARM ST 1	1
Abstract	The human-readable description of the individual record used to help determine the nature of the underlying physical data resource. Due to the general nature of the Catalog, a fair amount of information about the data resource may need to be captured into this one general element. Examples: Sample: Core Research Center, Cutting DD18210, from well operated by St. Michael Exploration, located in Weld County, CO, under lease 2-1 Grace State, with API number 0512310130. Sample: This is a geologic sample in one of the Bureau of Economic Geology's three Core Research Centers. API Number: 420513299400 Top Depth: 11744 Ft. Bottom Depth: 11767 Ft. sample_type_name: CORE CHIPS/CORE PLUGS sample_category_name: Core formation_name: Unknown formation_age_name: Unknown facility_name: Houston reservoir_name: BILOXI CREEK WILCOX operator_name: APACHE CORPORATION state_name: Texas county_name: Burleson	1
DataType <sup>2</sup>	A controlled vocabulary of data types. An item may include multiple dataTypes, including: 1) Auger Samples, 2) Fluid Samples, 3) Geochemical Samples, 4) Hand Samples, 5) Ice Cores, 6) Paleontological Samples, 7) Rock Cores, 8) Rock Cuttings, 9) Sediment Cores, 10) Sidewall Cores, 11) Thin Sections and Polished Sections, 12) Type Stratigraphic Sections.	1-N
SupplementalInformation	This standard field will be used to provide specific information on how to access the physical data represented by the metadata record. This may be general for the entire collection (e.g., a URL to another Web site) or a specific reference to an online resource like an ordering system with a specific ID. Example: Sample: Repository managed by the USGS Core Research Center, additional information can be found at <a href="http://geology.cr.usgs.gov/crc">http://geology.cr.usgs.gov/crc</a> ; Sample: Web (this sample): <a href="http://inet1.beg.utexas.edu/crc2/geosample.aspx?ID=160580">http://inet1.beg.utexas.edu/crc2/geosample.aspx?ID=160580</a> Phone: 512-471-0402 (Austin CRC) Phone: 713-466-8346 (Houston CRC)	1
Coordinates	Geographic longitude and latitude. Both values shall be contained in the same element and be listed in the order: longitude,latitude with values separated by a comma. Example: Sample: -118.023423, 45.02312	1
DatasetReferenceDate	A reference date indicating currency of the underlying data record, which may be the date the metadata record was assembled for the National Digital Catalog. Proper date formats are defined in ISO 8601, which include: 1) □yyyy, 2) yyyy-mm, 3) yyyyymmdd, 4) yyyy-mm-dd	1

## OPTIONAL METADATA

Required Element Name	Definition	# <sup>3</sup>
AlternateTitle	Collection owners may elect to provide additional title identifiers for individual records for further identification or use by other Web service interfaces. The AlternateTitle field may include either textual titles or specific sample IDs used by the collection.	1
AlternateGeometry	The underlying collection resource may use an alternate method of storing a geospatial footprint. If so, this text field should be used to describe the authoritative source for geographic location and how the simple coordinates were derived.	1
OnlineResource	One or more URL pointers to textual information about the specific record.	1
BrowseGraphic	One or more URL pointers to images representing the specific record.	1-N
Date	If a meaningful date within the geosciences domain can be attached to the record (e.g., a collection date), it can be supplied here. Either date may be to any degree of precision, or may be left blank to indicate uncertainty. Examples are 2001, 2001-03, 1939-1945, 20030331, 2000-03-31.	1-N
VerticalExtent	Vertical extent information can be provided and is especially useful for rock core samples. Specification of extent can contain three elements: minimum value, maximum value, and unit of measure. These elements will be collected as 2 or 3 values representing the UnitOfMeasure and MaximumValue with the possible addition of MinimumValue (e.g., m,35.4,0 for a rock core measured at 35.4 meters).	1

## 2 – Map Existing Metadata to National Digital Catalog Metadata

Having an understanding of the required and optional metadata elements accepted by the National Digital Catalog, you will now need to map the sample-level properties found in your own particular collection to these metadata elements.

In some cases this might be a one-to-one mapping where, for instance, the required “datasetReferenceDate” element might correspond to 'DateCreated' property in your own collection database. In other cases, properties from your own database may be used to populate a single metadata element. This could be the case for a required element such as “abstract” where it may be desirable to concatenate a number of fields to provide a richer description of the resource and more specific search results. In other cases, there may be a need to transform data from your database to a value suitable for the National Digital Catalog. For instance, if the location information were stored in township/range/section, it would be necessary to convert these values to a single geographic latitude and longitude point to populate the required coordinates metadata element.

### 3 – Select Upload Format

The National Digital Catalog supports two primary formats for loading data. The simplest format is commonly referred to as a Comma Separated Value or CSV format. This format is fairly easy to create and many software packages use this format to exchange tabular data. The more complex format accepted by the National Digital Catalog is a custom XML format that provides better handling for repeatable elements than possible with the CSV format. The XML format eliminates some parsing issues present in the CSV format. These formats are discussed in more detail below.

#### **CSV File Format**

A CSV file format represents tabular data. Each line in the file corresponds to a single record, and the properties associated with each record are delimited on the same line by a comma character. The comma is a special character that separates data fields, which is useful for recording record/sample characteristics. However, comma delimited files may present issues when the data field that is being delimited itself includes a comma, which requires enclosing the entire field in double quotes so that the embedded comma is not regarded as a field delimiter but as part of the field's information. This works well unless the information in the field also uses a double quote as well as a comma to indicate (for the sake of example) inches in a measurement. In general, the comma separated value doesn't work well for many types of data. Therefore, the National Digital Catalog allows use of another "record delimiter" or "pipe" character (|) to delimit fields rather than a comma. The pipe character is seldom used in normal text and allows fields to include the more common comma and quote characters without risking confusion with the field separator. For example, a single record with four fields might look like the following using the pipe record delimiter character:

```
1341234|This is a Title|m,35.4,0|Please note, the existence of a comma and "quote" characters.
```

The above line could not be represented accurately using a comma delimited approach but is fairly straightforward using the chosen record delimiter character.

CSV formatted files must include a single record on the first line that indicates the metadata element names and delimiters corresponding to the subsequent data (similar to a map legend). Applying this to the above example, the first two lines of the CSV file would look like the following:

```
COLLECTIONID|TITLE|VERTICALEXTENT|ABSTRACT  
1341234|This is a Title|m,35.4,0|Please note, the existence of commas and quote characters.
```

#### **XML File Format**

The XML file format is more complex than the CSV, can better handle repeating values for the same record, and is normally more reliable for handling various data inputs. It is suitable for those who are familiar with the XML format and have tools available to work directly with this format. When using the XML file format, new elements are added to the required and optional metadata elements listed above to handle repeatable elements. For instance, the alternateTitle element has one-to-many "title" child elements to handle multiple alternateTitle elements.

An online XML template is available at: <http://datapreservation.usgs.gov/docs/collectionMetadataExample.xml>

A basic two-record sample is available at: <http://datapreservation.usgs.gov/docs/NGGDPPSampleMetadata1.xml>

XML schema is provided for basic XML validation: <http://datapreservation.usgs.gov/docs/NGGDPPMetadata.xsd>

These examples outline structure of the XML file acceptable for upload to the National Digital Catalog.

## 4 – Extract Sample-Level Data into Upload Format

The extraction process transforms the records in your collection into National Digital Catalog metadata elements and stores this in the selected file format (CSV or XML). A repeatable process is desirable for subsequent updates.

Most databases may output CSV formatted files through built-in utilities, or customized queries. Microsoft Excel may be used to produce CSV formatted files but the options for character encoding and delimiters are limited. Excel uses the system settings to determine character encoding and delimiters. On a Windows platform these can be set by accessing the Control Panel --> Regional and Language Options. This option is not available on a Macintosh.

Some databases may produce information in an XML format, which may be not be acceptable by the National Digital Catalog but may be transformed using an additional XSL stylesheet into the correct format.

## 5 – Upload Metadata to National Digital Catalog

Once a correctly formatted extract file is available, it may be loaded into the National Digital Catalog through the provided web interface. This interface requires the data loader to login to the ScienceBase data management platform to add data to the desired collection(s). You will need a login account, which will include specific permissions allowing you to write to your State's digital repository in the National Digital Catalog. A login account may be requested by sending email to [myusgs@usgs.gov](mailto:myusgs@usgs.gov) with a request, similar to: I am a data steward for a NGGDPP geophysical/geological collection. I am requesting a myUSGS user account with NGGDPP\_Author role to upload sample metadata to the National Digital Catalog for "State Name". My contact information is

Agency, First Name, Last Name, Email Address, Phone Number, State for which you plan to submit metadata

After receiving a myUSGS user account, please follow data upload directions available at:

<http://datapreservation.usgs.gov/docs/NGGDPP/NGGDPPQuickGuide.pdf>

## Contact Information

Questions and concerns may be directed to:

Natalie Latysh, U.S. Geological Survey - Core Science Systems  
303-202-4637  
[nlatysh@usgs.gov](mailto:nlatysh@usgs.gov)

Betty M. Adrian, Acting National Geological and Geophysical Data Preservation Program Coordinator  
U.S. Geological Survey  
303-202-4828  
[badrian@usgs.gov](mailto:badrian@usgs.gov)