

Guidance for Checking Metadata with Data

Properly validating the metadata for a dataset involves two components:

- 1) Ensuring that the metadata record adheres to an established standard, such as FGDC-CSDGM or the ISO metadata standard. This document outlines instructions for using the FGDC-CSDGM standard.
- 2) Ensuring that the metadata record contains all the necessary information to accurately and adequately explain the information contained in the dataset that is being described.

Validating a Metadata Record in the FGDC-CSDGM Standard

The FGDC-CSDGM ‘standard’ is a formal set of requirements that detail the required elements of a metadata record, as well as the structure or order that they are required to follow. The simplest and most definitive way to check a record for compliance is with the USGS Metadata Parser (MP) tool.

The tool is freely available online here: <http://geology.usgs.gov/tools/metadata/tools/doc/mp.html>

Once downloaded, the executable version of the tool (the program file ending in “.exe”) can be run from the command line with several very simple commands. Please see the “Running MP from Command Line” document for additional details on how use this program to check a metadata record for compliance against the FGDC-CSDGM standard.

Another option to check a metadata file for compliance is to use the online validator tool that Peter Schweitzer (USGS), the creator of the MP tool, also developed. This tool is available online here: <http://geo-nsdi.er.usgs.gov/validation/>, and allows a user to upload a file through their internet browser to check for errors in structure and content.

Both of these methods will probably require some familiarity with the FGDC-CSDGM metadata standard. For the most part, the standard is not too difficult, but it is very extensive. As a fair warning, deciphering the errors that MP finds in a metadata record can be a bit cryptic for a new user, and a full explanation of possible errors is beyond the scope of this document. Although there is a learning curve, the best approach is usually to just begin getting familiar with the FGDC-CSDGM standard.

MP will identify when an element is found out of place, when elements are missing, or when they contain non-permitted values. A user will usually have to look at the errors that MP identifies and then open the XML file and consult the FGDC-CSDGM standard to update incorrect elements and values. An excellent way to read an XML metadata record and actively edit the file is to open it in XML Notepad with the FGDC-CSDGM schema applied, which are both free resources. See the “Using XML Notepad for FGDC” document for more information on this approach.

The full FGDC-CSDGM metadata standard workbook (that outlines proper formatting, structure, and permitted values) can be acquired here:

http://www.fgdc.gov/metadata/documents/workbook_0501_bmk.pdf

The FGDC-CSDGM standard is also presented in an interactive website here:

<http://www.fgdc.gov/csdgmgraphical/index.html>

Quality Assurance / Quality Control (QA/QC) for a Dataset and its Metadata

After a metadata record has been determined to be in compliance with the FGDC-CSDGM standard (i.e., it “passes” MP with no identified structural or content errors), it should ideally also be read as a complete document and viewed alongside the data that it actually describes. This is because while a metadata record may comply with basic structural and organizational tests, the record could still possibly contain inaccurate or unclear information or otherwise be missing necessary details.

The following is a checklist of components to consider when QA/QCing metadata alongside data:

Specific Metadata Elements

- *Abstract, Purpose, Supplemental Info, etc.*

Does the abstract or the general description of the data product make sense? The abstract or the description of the data set (perhaps captured in several different elements in the metadata record, including ‘Abstract’, ‘Purpose’, ‘Supplemental Info’, and ‘Use Constraints’) should collectively provide enough information for a reader to gain an adequate understanding of the information that is represented in the dataset. While more topical expertise might be necessary to properly use a scientific dataset for in-depth analysis, a general reader should be able to acquire a working knowledge of the information that is contained in a dataset from the metadata file alone. When reading an abstract as a reviewer, make sure the text makes sense to you and that you have solid grasp of what “the dataset” actually consists of (e.g., GIS dataset, a set of related tables, a stand-alone tabular file, a database of recordings, a web resource, etc.) If you have questions, chances are, others will too. If, after looking at the actual files or materials in a data collection, it appears that there either A) materials present in the data package that are not described in the metadata file, or B) files or data that are described in the metadata that do not appear to be present in the actual materials, it may be necessary to contact the data producer to clarify things or obtain missing resources.

- *Contact Information*

Needing to find additional information about a data product can be a common scenario for downstream data users. Valid contact information that will allow future users to get in touch with an individual or staff at the agency or organization that produced the data is critical. While this may not mean that it is necessary to dial every number or send test emails to each address found in a metadata record, make sure that phone numbers, addresses, and email addresses at

least appear to be complete, and if possible, accurate.

- *Any Online Links / Web Services*

A simple check to make sure that any online linkages to described resources, links to sites of larger projects associated with the data, online distribution points or web services are all valid can be a quick and very useful QA/QC measure. Given that digital distribution of datasets is usually the primary means of data distribution/acquisition, it is worth ensuring that a typo doesn't prevent would be users from accessing any described resources. These are worth checking—try navigating to the URL's listed.

- *Spatial Reference Information*

For any data products that could be considered geospatial data or that otherwise have a geospatial component, verifying that the spatial reference specified in the metadata record matches the information inherent in the data is a very valuable editorial step. Although most Geographic Information System (GIS) software will allow a user to view or retrieve this information from a file, metadata files can become separated from the data they describe. Ensuring that the spatial reference listed in the metadata file and the actual spatial reference match when a dataset is officially released will avoid confusion or uncertainty downstream as to whether a dataset has been potentially altered or compromised. Depending on the file format or distribution method, it may also be possible to lose the inherent spatial reference (i.e., a GIS file could be converted to a stand-alone table of data with geographic coordinates, etc.), in which case, spatial reference information can be absolutely integral to being able to use the data properly. UTM measurements in columns titled "X" and "Y" are hard to use properly without knowing which spatial reference was used to record them. Ensure that this information is included and correct whenever applicable.

- *Keywords*

Having at least one topical ('theme') keyword is required for an FGDC-CSDGM metadata record. However, having a more comprehensive list of keywords is very valuable for ensuring data catalogs and search functionality will be able to efficiently index or retrieve a metadata record. Making sure the keywords make sense, and possibly adding any additional relevant terms for 'theme' or 'location' may improve a metadata record. When appropriate, using a controlled set of keywords from one or more specific thesauri can also be valuable. While it may be difficult to do, checking to make sure that any keywords listed as being from a specific thesaurus are actually valid members of the listed thesaurus can be a valuable QA/QC step.

- *Entity and Attribute Information (Description of Column Headings and Contained Values in any Tables)*

This is one of the most important pieces of information that a metadata record can contain. In a nutshell, all of the tables, the columns they contain, and the values that are present in each row need to be clearly interpretable and documented. This should not be more laborious task than simply stating what tabular information is present with dataset, being sure to include the units that were used when applicable, and clear definitions for any coded value sets and/or abbreviations. This documentation is critical for downstream data users; without it, many datasets literally become useless.

When reviewing data alongside metadata, ensure that all the tables and columns in the actual dataset are described in the metadata record and also that all values present are consistent with the definitions and explanations that are provided. A logical check to make sure that the values present in the columns make sense as reasonable values for the field is an essential QA/QC measure (e.g., do temperature values fall within a realistic range?, do measurements make sense given the units and the scope of the project?, do all the values in a domain-controlled field have a definition?, etc.)

Assessing the Actual Dataset – A Quick Checklist

- Open the dataset up in the software in which it is intended to be used. Does everything display properly? Are there any undocumented version issues or use notes that future users should be aware of that are not captured in the metadata?
- Is there anything about the size of the data product or distribution protocol that affects how future users might interact with it? It is good to package data products and documentation (metadata) as manageable bundles that can be managed and distributed jointly.
- When opened up or viewed, does the dataset or product generally seem consistent with the description that is provided in the metadata record? Many issues can be detected during a simple check to make sure nothing was overlooked and that final re-naming or file organization at the end of a project did not introduce any problems.
- Take a moment to step back from the data product and make sure that the organization and the materials would make sense to someone after reading the provided documentation. If things are complicated or unclear to you, there is a high chance they will be to someone else as well. A relatively small amount of time spent organizing and documenting data at the end of large project can make a critical difference in the value it can serve to the research community if it ensures that the data will be used properly and/or made more accessible in the future.

*"U.S. Geological Survey - Data Management Guidance Materials / White Papers" (Various Authors)
Resources obtained from USGS staff and/or the USGS Data Management website
(<https://www.usgs.gov/datamanagement>), November 2015. Documents may be subject to revision.*