How to review metadata

by Peter Schweitzer

First, you'll want to get the record into good shape structurally. Second, review some specific groups of elements that people tend to have trouble with. Third, generate FAQ-style HTML output using the USGS metadata parser tool and read it as though you were a non-expert user, looking for unanswered questions.

1. **Check and fix the structure**

   Run the metadata parser (mp). The web version of mp is handy for first attempts. If you're using the downloaded version, generate an error file using the `-e` switch (`mp input_file.met -e err`). There are two ways to look at the error file. You can just read it or you can run `err2html` to generate a more friendly-looking report. The things to fix first are "ambiguous indentation" and "Extraneous text following..." which indicate hard-to-spot indentation problems. Next focus on any messages that say "no element recognized", then on any that say "is not permitted in". These are the most important structural problems and all need to be fixed. Then look carefully wherever you get a "too many" error; these can almost always be done better.

   Don't worry too much about missing or empty elements or improper values at this stage of the review. Deal with those later.

2. **Specific elements to watch**

   Identification_Information

   Title

   This should be informative. It's going to show up as the title of the web page that mp generates, and it will be the only thing that shows up if this record is returned from a search of the Clearinghouse or web. The coverage name by itself is not enough. If you have a lot of records from the same publication, give them a colonized title, "Main title: subtitle" where the main title is the same for all of the records.

   Geospatial_Data_Presentation_Form

   Most of the time this is "map". If it's imagery, you can use "remote-sensing image". Any text is permitted, but it's best to be consistent with what other people have written before.

   Series_Information

   I try to use the same text for all of the series names, and to write the Issue_Identification in the same way. For Series_Name I use forms like these

```
U.S. Geological Survey data release

U.S. Geological Survey Data Series

U.S. Geological Survey Open-File Report

U.S. Geological Survey Scientific Investigations Report

U.S. Geological Survey Professional Paper
```

for Issue_Identification, you can use MF-xxxx and I-xxxx and DDS-xxx but for OFRs and Professional Papers, it's better to use only the numbers, not OF or PP.

Supplemental_Information

The old AML DOCUMENT used to put stuff in here that doesn't belong. You'll know that's what you have if you see headings that aren't real FGDC elements:

```
Procedures Used

Reviews_Applied_to_Data

Related_Spatial_and_Tabular_Data_Sets

Other_References_Cited
```

Most of the information in these "subsections" really goes into Process_Description, Cross_Reference, or Source_Information.

Browse_Graphic_File_Name

Convert to a URL if you can. That way people can get it easily.

Point_of_Contact

This is not strictly required by the Standard, but it really should be. Don't accept metadata without it.

Data_Quality_Information

Attribute_Accuracy_Report

Frequently what's written here is meaningless. If it doesn't give you any useful information, delete it.

What really goes here is how the authors checked the attribute data. Certainly if you know there are errors and you aren't going to fix them before release, write that up too. But that rarely happens; most of the time we think we have it right, so here's the place to say what we did to review them.

Logical_Consistency_Report

Often only a useless statement about topology. To be more than that, authors should construe this as indicating whether the geographic features and their attributes have the same meanings in different parts of the map or data set. Was the processing done by different people? Are the mapped units divided uniformly throughout the extent? State-line faults are a good example of something you would write up here.

Completeness_Report

What's missing, in geography, time, or attributes? Don't just say "complete".

Source_Information

Frequently problematic. Check to see that each Source_Citation is within its own Source_Information. Ignore errors about missing Source_Time_Period_of_Content unless the source is time-sensitive data. Make up a Source_Citation_Abbreviation if one is not given; this should be something like "Smith 1997" and its value will show up in a Source_Used_Citation_Abbreviation of a Process_Step if those are done right. Comments above for Series_Information apply to sources as well. Source_Contribution should be brief but informative; what did the authors get from this source? There's more information on these relationships in Metadata in Plain Language, at <https://www.usgs.gov/products/data-and-tools/data-management/metadata>

A common question is whether to include as sources all of the references given in the report. I would include only those references from which data were taken directly, so for example if you can point to a line or attribute value and say "that came from Smith's 1997 map", make Smith's 1997 map a source. If you'd like to include other sources, they can be placed in the Cross_Reference section.

Process_Step

If you have the exact date that the Process_Step was completed, use that. While the exact date is often difficult to specify, it is required by the Standard. Don't sweat it too hard, but do try to come up with a reasonable date if you can. Unfortunately the Standard does not allow you to give this as either an approximate date or as a range of dates. That's a weakness of

the Standard. Use Process_Contact only if that person is not the Point_of_Contact and especially if the person is not listed as an Originator.

Entity_and_Attribute_Information

This section is often difficult for authors. It shouldn't be, but it takes some time and attention to detail. Here are some ideas to help.

Overview_Description

These typically don't contain enough information for a real end-user to understand the data effectively. It's simply not enough to do `ITEMS cover.PAT` and paste that into the file. People need to know what the field names mean in real terms, what the values mean individually if they are abbreviations, and what the units of the numbers are. Also people need to know what value is used to indicate missing data. All of these are better expressed in a Detailed_Description. As a reviewer, you might not be able to persuade the author to do a Detailed_Description, but you should insist that the information that real data users need be there.

Detailed_Description

Entity_Type

Use the ARC/INFO table name if it's a coverage, or the DBF name if it's a shapefile.

Attribute

Don't document the attributes that ARC/INFO creates for its own use. So omit AREA, PERIMETER, LENGTH, FNODE#, TNODE#, LPOLY#, RPOLY#, cover#, and cover-ID. Document only things that actually have some scientific content added by the authors.

If their value would be "author" or "this report", omit both

Attribute_Definition_Source and

Enumerated_Domain_Value_Definition_Source

even though mp will flag it as an error.

Read Metadata in Plain Language for help on how to do *Attribute_Domain_Values*. These are often done wrong, but doing them right means you end up checking the values, so it's a really good idea to do them right.

Distribution_Information

If there's no Standard_Order_Process, people can't get the data. So make sure that one is present. mp doesn't flag it as missing because it's mandatory if applicable. But it's not optional.

Digital_Transfer_Information

Put something into *Format_Information_Content*. mp uses this in its FAQ-style HTML. It should be a plain-language statement about what data are conveyed using this format.

Format_Name is the data format. Common values are

- `Arc/Info Export format`
- `ASCII`
- `Excel spreadsheet`
- `Tab-delimited text`
- `SDTS`

FGDC offers this list of values for Format_Name.

Format_Version_Number is

- The ArcGIS version number if it's an export file
- `1.0` if it's a shapefile
- whatever Excel version if it's a .xlsx
- skip if it's tab-delimited or comma-delimited text

Format_Version_Date should not be used unless a date is how that particular format is distinguished from other formats. This rarely occurs.

Format_Specification should be skipped unless you're dealing with a non-standard format, in which case describe it in detail here.

Transfer_Size is in megabytes. Use only the numeric value for size.

What formats and files should be documented in this way? I prefer to see the main data files or packages of them done. Often authors will make individual files available for download, and will include PDF's or PostScript versions as well, and additional text. The metadata needs to focus on the data, but it's okay to describe how to get these ancillary files too.

Metadata_Contact

This should be the person who wrote the metadata. It should not be an author whose anonymous technician wrote the metadata, especially if the author doesn't know anything about it. If the author's lab rat wrote the metadata, and the lab rat has departed for a real job somewhere else, just change the contact email address and phone to point to somebody else. But it helps to know who really wrote the metadata.

Go back through the record and remove any empty elements. Tkme can do this for you if you choose Prune from the Edit menu.

Read-through

Run `mp` and generate FAQ-style HTML. Do this with the `-f` switch, like this: `C:> mp myfile.met -e myfile.err -f myfile.faq.html`

Now look at myfile.faq.html with a web browser. Look for questions that don't have answers, and ask whether there really should be an answer for these. Check to see whether the links work. Read the answers to see whether anything comes out strangely.