Ten Most Common Metadata Errors

From the FGDC and the National Metadata Cadré

Metadata experts have joined together to compile this list of common errors that are typically made when people create metadata records. If we learn from their experience, we can all create more reliable and accurate metadata. Review the list and note which ones are on your personal list of errors. The next time you document your metadata, consider reviewing the list again so that you can create a more complete and accurate set of metadata records.

Error # 10: Defining your data set too finely or too broadly

It's easy to become overwhelmed trying to individually document every data table and resource. On the other hand, trying to cover all of your data resources with a single metadata record will drive both you and your data users crazy. A good rule of thumb is to consider how the data resource is used — as a component of a broader data set or as a stand-alone product that may be mixed and matched with a range of other data resources.

Error # 9: Using incorrect State Plane Coordinate System Zone Identifier values

The default SPCS Zone Identifier (4.1.2.2.4.1) values for some software products are based on early BLM values rather than the FIPS Code prescribed by the CSDGM.

Error # 8: Confusing "Currentness Reference" with "Publication Date"

While the Currentness Reference (1.3.1) may refer to a publication date, it is actually a qualifier to Time Period of Content (1.3).

- ❑ Does the time period refer to the date and time of *data capture* or *ground condition* — as in photography or field data collection?
- ❑ Does it refer to the date the information was officially *recorded* — as in a deed?
- ❑ Does it refer to a *publication* date — as in a "1978 USGS Topo map"?

Basically, the idea is to let prospective users know how well you are able to "nail" the actual time period of content.

Don't Duck Metadata

Error # 7: ***Misunderstanding resolution***

Who could blame us? The purpose of these fields is to indicate how coarsely or finely information was recorded. For example:

- Latitude Resolution (4.1.1.1) and Longitude Resolution (4.1.1.2)

  These values represent the minimum possible difference between coordinate values. For example:

	resolution (4.1.1.1 or 2)		geographic coordinate units (4.1.1.3)
30° 30’ 30”	0.0028	(1° / 360”)	degrees, minutes, seconds
30° 30’ 30.01”	0.000028	(1° / 36,000”)	degrees, minutes, decimal seconds
30.00001°	0.00001	(1° / 100,000)	decimal degrees

- Abscissa/Ordinate Resolution (4.1.2.4.2.1 and 2)

  These values represent the minimum difference between X (abscissa) and Y (ordinate) values in the planar data set.

  — For raster data, the values normally equal pixel size; e.g., 30 (TM).

  — For vector data, the values usually indicate the “fuzzy tolerance” or “clustering” setting that establishes the minimum distance at which two points will NOT be automatically converged by the data collection device (digitizer, GPS, etc.).

  Note: Units of measures are provided under element Planar Distance Units (4.1.2.4.4) and would be “meters” for the TM example provided and likely millimeters for the vector example.

Error # 6: ***Putting too much faith in metadata tools***

Human review is the only thing that matters. The tools are there to help you, but they can’t do the job alone. Remember: *Garbage In — Garbage Out*.

Error # 5: ***Taking the minimalist approach***

A common overreaction to the expansive nature of the CSDGM is to adopt “minimal compliance” as an operational approach. Limiting your documentation to the *required* portions of Sections 1 and 7, or even to all *required* fields, will limit the value of your effort and the metadata records you produce.

Instead, identify those fields that apply to your organization and data. Then create functional templates, or subsets, of the CSDGM.

Don’t Duck Metadata

Error # 4: ***Understanding assessments of consistency, accuracy, completeness, and precision***

Section 2. Data Quality Information is intended to provide a general assessment of the quality of the data set. This represents the Achilles heel for many RS/GIS professionals. Consider it an opportunity to get to know your data set.

A brief summary:

- Attribute Accuracy Report (2.1.1) — Assessments as to how "true" the attribute values may be. This may refer to field checks, cross-referencing, statistical analyses, parallel independent measures, etc. *Note:* This does NOT refer to the positional accuracy of the value (see 2.4).

- Logical Consistency Report (2.2) — Assessments relative to the fidelity of the line work, attributes and/or relationships. This would include topological checks, arc/node structures that do not easily translate, and database QA/QC routines, such as:
  - Are the values in column X always between "0" and "100?"
  - Are only text values provided in column Y?
  - For any given record, does the value in column R equal the difference between the values provided in columns R and S?

- Completeness Report (2.3) — Identification of data omitted from the data set that might normally be expected, as well as the reason for the exclusion. This may include:
  - Geographic exclusions (*Data were not available for Smith County.*);
  - Categorical exclusions (*Municipalities with populations under 2,500 were not included in the study.*); and
  - Definitions used (*Floating marsh was mapped as land.*)

- Positional Accuracy (2.4) — Assessments of horizontal and/or vertical positional (coordinate) values. Commonly includes information about digitizing (RMS error), surveying techniques, GPS triangulations, image processing or photogrammetric methods.

- "Precision" — An indication as to how "finely" your data were recorded, such as digitizing using single or double precision. *Note*: The precision of the value in no way reflects its accuracy or truthfulness.

Don't Duck Metadata

Error # 3: ***Glossing over Section 5. Entity and Attributes***

Another of the GIS professional's "Achilles tendons," this section maps out data content and should be a product of your data design effort.

- Use the relational database format as a *guide*:
  Entity Label (5.1.1.1) – Table Title
  Attribute Label (5.1.2.1) – Column Titles
  Attribute Domain Values (5.1.2.4.X) – Recorded values within each column

- Domain Types – set of possible data values of an attribute

  Enumerated Domain (5.1.2.4.1)
  — A defined pick list of values
  — Typically categorical, such as road types, departments, tree types, etc.

  Range Domain (5.1.2.4.2)
  — A continuum of values with a fixed minimum and maximum value
  — Typically a numeric measure or count, may be alphabetic (A–ZZZ)

  Codeset Domain (5.1.2.4.3)
  — A defined set of representational values
  — Coding schemes, such as FIPS County Codes or Course No. (GEOG 1101)

  Unrepresentable Domain (5.1.2.4.4)
  — An undefined list of values or values that cannot be prescribed
  — Typically text fields, such as individual and place names

- Entity Attribute Overview (5.2.1)

  A summary overview of the entities/attributes *as* outlined in *either* Detailed Description (5.1) *or* an existing detailed description cited in Entity Attribute Detail Citation (5.2.2). *Note*: The field should not be used as a stand-alone general description.

Error # 2: ***Thinking of metadata as something you do at the end of the data development process***

Don't Duck Metadata

Metadata should be recorded throughout the life of a data set, from planning (entities and attributes), to digitizing (abscissa/ordinate resolution), to analysis (processing history), through publication (publication date).

Organizations are encouraged to develop operational procedures that

- institutionalize metadata production and maintenance, and
- make metadata a key component of their data development and management process.

Error # 1: ***Not doing it!***

If you think the cost of metadata production is too high, you haven't compiled the costs of **NOT** creating metadata:

- loss of information with staff changes,
- data redundancy,
- data conflicts,
- liability,
- misapplications, and
- decisions based on poorly documented data.

This is a product of the Federal Geographic Data Committee (FGDC) Metadata Education Program and the National Metadata Cadré. Much of the information was collected during metadata workshops, meetings, and presentations and represents the comments and contributions of many individuals.

Specific contributors include:

Lynda Wayne, editor, FGDC contractor • Anne Ball, NOAA, Coastal Services Center • John Bocchino, New Jersey Department of Environmental Protection • Chris Cialek, Minnesota State Planning • George Lienkaemper, USGS, Forest and Rangeland Ecosystem Science Center • Jackie Olson, USGS, Water Resources Division • Richard Pearsall, FGDC • Hugh Phillips, 3001, Inc. • Nina Savar, Northeastern Illinois Planning Commission • Peter Schweitzer, USGS, Geologic Division • Sheryl K. Soborowski, Virginia Tech • Bruce Westcott, RTSe, Inc.

Don't Duck Metadata

What are the top ten most common metadata errors?

Responses from the experts

***Nina Savar**, Northeastern Illinois Planning Commission*

What are my top 10 most common metadata errors? First, I'll reinterpret the question as: *What are the top 10 most confusing metadata fields* and go from there. (The following are in no particular order of importance.) I'll add a thought or two about the metadata process, as well.

10) What is resolution (and why it's a critical field to fill in); i.e., abscissa/ordinate resolution, latitude/longitude resolution, depth resolution?

9) Untangling the distinctions between "Access Constraints," "Use Constraints," and "Distribution Liability."

8) Does "Logical Consistency" include more than discussions about graphic integrity?

7) A table (i.e., cookbook) that associates all the "SDTS Point and Vector Object Types" to actual vendor terminology. For example: Is an ArcInfo line coverage a *network chain*, *planar graph* or a *complete chain*? Is an AutoCad Drawing file a *string*? What is an ArcView Shape file polygon? What's the difference between a *point* and *entity point*? I think this is critically needed.

6) Understanding that "Contact Information" is used differently in each of four sections, that is:
   Section 1: the person who knows the most about the data set.
   Section 2: the person(s) who gathered the data and/or put the data set together.
   Section 6: the person who can distribute the data.
   Section 7: the person who wrote the metadata.

5) Understanding all the fields required for defining offline media like recording density, units, and format. Could a table of the most common media types (by brand) be developed? I had a bear of a time and still think my fields are incorrectly populated.

4) For ESRI users who store their data in a State Plane Coordinate System, they need to know that the numbers ESRI uses are not the same FIPS codes that are requested by the CSDGM, but who would know that?

3) Need some help or examples of how to describe spatial data accuracy and attribute accurac

2) I would like to see the following: Wherever the *Green Book* or updated CSDGM workbooks reference other documents (such as FIPS documents containing definitions), they should be put online somewhere (if they are not already) and the url to their locations should be included. I would expand reference materials to include some standard thesauri for place names or themes and make them digital and/or reference their urls.

Don't Duck Metadata

1) Making sure that folks who rely on mechanical means of populating their metadata fields (by using a tool or script) know that they will not necessarily create useful metadata. I really believe that metadata needs to be written and reviewed by humans before it is useful. I think it's important to move away from the notion that metadata is put off as the last thing to do, but use it instead to propel a project and elevate metadata an integral part of our work.

John Bocchino*, *New Jersey Department of Environmental Protection

On the top ten topics in no particular order:

- ❑ currentness reference
- ❑ attribute accuracy report
- ❑ logical consistency report
- ❑ horizontal/vertical accuracy report
- ❑ proper/accurate documentation of lineage/source information
- ❑ SDTS point and vector object type
- ❑ Section 4: Spatial Reference Information (Abscissa/ordinate, lat/long
- ❑ resolution, etc.)
- ❑ proper/accurate documentation of Entity and Attribute information
- ❑ network resource name vs. Online linkage
- ❑ understanding the structure/hierarchy of the CSDGM and the use of templates
- ❑ Sections 8-10 in the standard

However, the most important thing that must be conveyed is the *importance* and *relevance* to create compliant metadata. The user most be convinced that this is important and the time, effort, and money involved will pay off in the end and thus save time, effort, and money.

Peter Schweitzer*, *USGS, Geological Division (mp creator)

I would enlarge the problem beyond the metadata itself to the information processing in general and the process of data management. Here's my list, Letterman-style:

10. (*for Arc/Info users*) Taking time to document things that are consequences of the GIS, like making detailed descriptions of AREA, PERIMETER, LPOLY#, RPOLY# FNODE#, TNODE#, cover#, cover-ID, and the like. And (*for Arc/Info users*) simply dumping the results of ITEMS into an Entity_and_Attribute_Overview and calling that enough. People need to know the units of measured variables, and *percent* is _not_ a unit of measure.

9. Putting too much faith in mp. Human review is the thing that really matters. mp can help, but isn't the arbiter of what is and what is not good metadata. Prioritize errors like this, from most serious (fix) to least serious (understand and let go):

- Indentation problems
- Unrecognized elements
- Misplaced elements
- Too many of some element
- Missing elements
- Empty elements
- Improper element values
- Warnings and upgrades

8. Making too many metadata records. People who try to document every GIS coverage or data table can wear themselves out. Some aggregation is good for both the producer and the user. Ancillary coverages can be described as Source_Information.

7. Not making enough metadata records. Trying to cram all of the information about an entire research program into a single metadata record will drive you and your potential users crazy. Split when sources, processing, or spatial reference varies.

6. Agonizing over the most difficult elements. These include, but probably aren't limited to:

   Latitude_Resolution
   Longitude_Resolution
   Abscissa_Resolution
   Ordinate_Resolution
   Entity_Type_Definition_Source
   Attribute_Definition_Source
   Enumerated_Domain_Value_Definition_Source

   and, to a lesser extent,
   Attribute_Accuracy_Report
   Logical_Consistency_Report

5. Misunderstanding Enumerated_Domain as numerical values. Attributes measure (or count), categorize, or characterize real things. Those functions are expressed as Range_Domain (for measures or counts) and Enumerated_Domain (for categories). The hardest part is how to describe attributes whose values are simply text describing something, or are names like place names. This is a deficiency in the FGDC standard; there should be another type of Attribute_Domain_Values for descriptions; a better alternative might be to make Attribute_Domain_Values mandatory if applicable and not applicable if the values are sufficiently descriptive.

4. Substituting statements about precision for statements about accuracy. I do this often, because what I know is how variable the values are, and I don't know the true values that they estimate.

3. Larding the metadata with uninformative values. People can honestly disagree about this, but I find it aggravating to see more than a few *N/A*, *unknown*, *not applicable*, *implied*, or *see above* (or *below*). Reasonable defaults should be assumed. For example, if no Process_Contact is indicated, people should assume that the Point_of_Contact either

did the processing or knows who did, and that the people who did are either Originators or are listed in Data_Set_Credit. Likewise, any of the elements:

Entity_Type_Definition_Source
Attribute_Definition_Source
Enumerated_Domain_Value_Definition_Source

if missing, should be assumed to have the value *this report* or some similar self-reference.

2. Choosing a tool because it's free or because it's commercial. Making, maintaining, reviewing, and reading metadata cost so much more time and energy than the tools do that price per se shouldn't direct the choice of tools.

1. Not recognizing that "the metadata problem" involves not only tools and training, but also work-flow strategy and even the philosophy governing how your organization interacts with the users of its data.

\+ Not asking for help from the community. Beyond all the hype and promises (empty and fulfilled), beyond all the tools, training, and technology, what NSDI has done is bring a common language and common purpose to a highly diverse group of people, and we have found in each other consolation, challenge, and care.

George Lienkaemper, USGS, Forest and Rangeland Ecosystem Science Center

This list is a combination of user misunderstanding or lowballing and my pet peeves with the standard.

10. Abscissa and Ordinate resolution — without knowing digitizer specs, how do you know?
9. Process Steps — using the ArcInfo Log file
8. Radiocarbon dates — Thanks, Biological Profile
7. Geospatial data presentation form — looks like the paper world
6. Unrepresentable Domain — used too often
5. Currentness Reference — what?
4. Attribute Accuracy Report (or is it Logical Consistency?)
3. Logical Consistency Report (or is it Attribute Accuracy?)
2. Compliant Records (Sections 1 and 7 only) — if you've got data you've got data quality
1. Entity and Attribute Overview — for data users, it's the most important section. Why not do it right?

Don't Duck Metadata

Jackie Olson, USGS, Water Resources Division

Good question. I think one problem we see is in thinking of metadata as something that is done at the very end of the product cycle. Because it's a component that is often seen as a large chore and an extra piece of paperwork, it's relegated to back-end status.

One alternative is to approach it as something that is developed from the start of the project parallel with the data, as an integral part of the data in a simple record of information as it becomes available. There is more flexibility in input, time-savings, and a better record of the process.

Just about any record-keeping method the author prefers is useful for storing pieces of information; a simple text file, email to oneself, word-processing or database software, or even the metadata template.

Components can then be cut-and-paste entries to the final document. We often have the information already in another format and just need to adapt it for the metadata. Project/data information is presented in many existing documents that could serve as preliminary entries for fields such as the purpose, source and processing steps, accuracy, etc.

Project proposals, progress reports (quarterly, final or other), correspondence to or from other people about the data, and presentations, among others are documents we have to fill out anyway, so there's a time-saving benefit in recycling parts to metadata. The metadata entries are then updated as the data go through development.

Recording the entries as you go is also an advantage in that you're not trying to recall a lot of details maybe months afterwards. The record can be stored for access by all project personnel; it also serves as a history of ideas or steps that didn't work and why, as a reference tool for future data projects.

Bruce Westcott, RTSe USA, Inc. – SMMS Product Manager

The biggest error is the failure to adopt an operational plan that results in a sustainable program of creating, updating, disseminating and using metadata. My opinion is that the hurdles to this problem/error are much more important than errors that relate to the technicalities of the metadata content and the use of tools.

Chris Cialek, Land Management Information Center, State of Minnesota

Not appreciating its importance.

Don't Duck Metadata

Sheryl K. Soborowski, Fish and Wildlife Information Exchange, Virginia Tech

- Many agencies are experiencing trouble motivating personnel to take the time to do metadata. They feel it is a long and confusing process. Many comments were voiced that the FGDC Content Standard is difficult to understand. Even after reading a section's definition, the user would still not understand what information is being asked for. There

is confusion about what sections are absolutely necessary for a biological dataset. Some personnel were requesting that there be a separate standard for biological data that would exclude GIS specific sections. This was suggested in the hope that it would make the standard less daunting.

- ❑ Personnel and mid-level managers need creative ideas on how to justify allocating time and money to producing metadata.
- ❑ There was confusion expressed on performing clearinghouse searches. The agencies felt that their metadata was not accessible enough if the user did not know which clearinghouse node to search. Also, too many unrelated hits were returned.

***Lynda Wayne**, FGDC Metadata Education Coordinator*

- ❑ SDTS Point and Vector Object Types – how does one respond without the vocabulary?
- ❑ Misapplying supporting sections – contact, time, citation
- ❑ Not using the *Green Book* as a resource
- ❑ Using BLMt State Plane Coordinate Zone values instead of FIPS
- ❑ Misunderstanding *currentness reference*
- ❑ Using *Entity/Attribute Overview* as a general description option to fully documenting your data scheme. *Entity/Attribute Overview* should be used to reference existing, available data dictionaries and other forms of dbase documentation.
- ❑ Thinking that the metadata production is too costly
- ❑ Inability to effectively discriminate the *data set*. Either doing feature-level metadata or glomming independent data sets into a single non-specific metadata record.
- ❑ Not developing an effective organizational template
- ❑ Producing metadata at the end of the data development process instead of incorporating within your development process.

***Hugh Phillips**, 3001, Inc.*

- ❑ **Failing to include a browse graphic** — An important component of a GIS is the ability to view the spatial distribution of the objects that make it up — its object density and its relationship to other themes. If a browse graphic is created from the data set superimposed over a reference data set, such as a state outline, it provides spatial context to the browse graphic and the data set being documented. To some extent, a browse graphic creates a much more intuitive feel of the area the data set covers than does either the bounding coordinates or a place keyword. With a little thought, one can also create browse graphics for tabular data sets with indirect spatial reference (such as a county name), using another theme (such as a county boundary data set) as the canvas. Creating the browse graphic for a data set is one of the few fun parts to creating metadata, so don't skip it!

Don't Duck Metadata

- **Stuffing Distribution Liability into Access Constraints and Use Constraints**—The heading says it. In the zest to CYA, (and probably also for failing to understand the difference between these elements), some metadata creators intermingle the contents of these elements, even to the point that the values of the elements are identical.

- **Failing to write a very informative title** — The title of the data set is what is returned in a search of the NSDI. If it is fairly generic, say 'Tax Parcels,' the potential user may not be able to discriminate it from others that are also similarly generic. It is beneficial to richly load a dataset title. The title in itself can be mini-metadata. Consider this example: *North Atlantic Hurricane 'Tracks' 1886-1996 from NOAA NHC source data, Geographic NAD83, LOSCO (1998) [athgeog3dlnoa].* Encoded in this behemoth is a standard sort of title (which in this case also happens to include Time Period information and the producer of the raw source data), the coordinate system, the Originator, the Publication Date, and the systematic name of the dataset. Not all of this information will always be so handily available (or applicable) to insert into a 13% butterfat title and in other cases, one may find other useful information to include (such as the scale of the data set).

  If you were paying attention here, you see I just suggested that "combining of elements" that I complained about in the last point (Stuffing Distribution Liability…), and in spades, too. I think the Title element is sufficiently different and important to break the rule in order to increase its information content.

- **Not listing all the components of the metadata and data** — A data set and its metadata (inseparable really, the metadata is part of a complete dataset — I mean the GIS data files and the metadata that describe them— frequently consist of multiple files. A "shapefile" consists of a minimum of three files, but may have additional index files that aid the performance of the dataset or a projection file that defines its coordinate system. Metadata may consist of multiple files including ASCII metadata, HTML metadata, browse graphics, associated word processor documents, or list files. It is helpful to the data set end user to have an explicit listing of all these components of the data set (GIS data files and the metadata) in the metadata, that way they know exactly what they should have received as part of the dataset distribution package.

  This explicit listing is a standard component of Supplemental Information in data sets that I document. A distribution policy that packages the data with the metadata into a zip archive file is a good way to insure that the end user gets the whole schmear.

- **Range domain** — The Range Domain Maximum and Range Domain Minimum as defined in these elements is easy to complete if you have an attribute with an identifiable range. For an attribute of latitude in degrees, there would be no argument about a Minimum = -90 and Maximum = +90. It is very common to have numeric "range type" attributes that do not have such a well defined range; e.g., What is the Range Domain Maximum for a parcel's assessed value, $2 million? What is the Range Domain Maximum or Range Domain Minimum for the area of wetland polygons in acres? Lacking a real "domain," one may resort to determining the actual minimum and maximum values that occur in the attribute (which has a certain practical usefulness), but this is counter to the element definition and creates a currency problem if the data set is modified. Perhaps the best solution of all for a situation like the assessed value attribute mentioned above is to avoid the Attribute Domain Values subtree entirely and just write a meaningful Attribute Definition.

Don't Duck Metadata

- **Not using good English in narrative metadata elements** — Creating metadata is painful, but don't make the people who have to read your metadata suffer from incomplete sentences, telegraphese, undefined acronyms, and rampant spelling errors. The Abstract is the very essence of the description of a data set and probably the single most useful element in a metadata document. It is worth some extra time to craft its wording. Some would argue that the content and the intent of the abstract is much more important than its wording. That is true, but the user is much more likely to gain the correct and full value from the abstract if it is readily understood. One should also consider the impressions about "care" (in data set preparation) and potentially "quality" (of the data set) that is subtly communicated via its metadata. If the metadata for a data set are carelessly prepared, potential users can't help but be suspicious about the data set itself.

Don't Duck Metadata