Talk:Attribute Convention for Data Discovery 1-2 Working

From Federation of Earth Science Information Partners

Contents

List of Open Issues

You may add to this list (each issue gets a row).

Issue number	Issue name	Description	Reference below
1	Roles in Suggested section	Cleanup requested; current selection of role_entity not satisfactory	
2	Attributes that are part of NUG or CF	Identify which, if any, terms on our list are actually defined by another standard	

3	Guidance	How do we include/reference guidance?	Discussion (http://wiki.esipfed.org/index.php/Talk:Attribute_Convention_for_Data_Discovery_(ACDD)_W
4	Undecided Terms	Resolve open issues with terms	Resolved (review) (http://wiki.esipfed.org/index.php/Talk:Attribute_Convention_for_Data_Discovery_(ACDD)_Working#Attr , Open (discuss) (http://wiki.esipfed.org/index.php/Talk:Attribute_Convention_for_Data_Discovery_(ACDD)_Working#Attr
5	Purpose of document	Is this document (standard) just for discovery? a LOT of terms are clearly not discovery	
6	Internally complete	Does this document (standard) need to be internally complete per CF philosophy?	
7	Conventions attribute	NUG recommends putting all conventions into this single attribute; ACDD originally used Metadata_Conventions attribute. Do we suggest ACDD_1.2 for Conventions?	
8	Data type	Is the term cdm_data_type appropriate? We mean the THREDDS scientific layer; what terms are allowed?	
9	creator_institution duplicate	does creator_institution duplicate CF's 'institution'?	
10	metadata_link	NetCDF files should be self-documenting; this gives data writers an out. Also, it's not machine-usable, since the contents of the linked page could be ... anything	
11	geospatial_vertical_min/max	The current description seems to require the values to represent the distance from earth's center	Depth (http://wiki.esipfed.org/index.php/Talk:Attribute_Convention_for_Data_Discovery_(ACDD)_Workin _Graybeal_.28talk.29_19:17.2C_20_May_2013_.28MDT.29)
12	spatial and temporal bounds	should ACDD address these, given they can be misleading	Spatial and Temporal Bounds (http://wiki.esipfed.org/index.php/Talk:Attribute_Convention_for_Data_Discovery_(ACDD)_Working#Spa
13	date fields	Are the date_modified, date_created, date_issued fields deprecated in favor of the following elements date_content_modified,	Deprecated Date Fields (http://wiki.esipfed.org/index.php/Talk:Attribute_Convention_for_Data_Discovery_(ACDD)_Working#Dep

		date_values_modified, date_product_generated fields? Are those really the names we want?	
#	name	descrip	reference

Spatial and Temporal Bounds

There was extensive discussion on the list that reflected some deep philosophical differences, though I think they were more about priorities than disagreement about the principles. I summarized the status in the first quoted email below; then after some detailed discussions with Steve Hankin, wrote some guidance material as reflected in the second email below.

While this does not fully resolve the issue -- in particular, the recommended changes in B and C will be slow coming -- it is hoped that the issue is sufficiently addressed to move forward with approval of the document.

Per a suggestion, the section name Maintenance of Metadata in Derived Products was shortened to Maintenance of Metadata.

From: John Graybeal <john.graybeal@marinexplore.com> Date: April 24, 2014 at 2:04:06 PM PDT

```
Subject: Re: [Esip-documentation] Let's get rid of spatial and temporal bounds in ACDD

So this thread identified recommendations on several fronts for which we may have consensus:
A) Label risks when using global geospatiotemporal attributes (text to be added to TBD guidance;
needs to cover more than just geospatiotemporal attributes)
B) Improve server-side handling of computable dynamic attributes like global
geospatiotemporal bounds (recommend changes to server software)
C) Document and improve utilities' handling of global attribute creation
(recommend changes to utilities that create such files)

I think all these are mutually reinforcing, and see some useful next steps:
a) Propose text for
   a.1) the CF standard that provides guidance about attributes for file creators, updaters, and users.
   a.2) the ACDD geospatiotemporal attributes that says if they exist, their validity should be confirmed before use.
b) Make change requests to authors of servers describing desired strategies for maintaining and serving current attributes.
c) Define current practices of utilities and recommend changes, for example by making change requests.

If there aren't major objections, I'm willing to tackle (a); and I have already created a tracking page for
(c)[1] but could use some more authoritative help. (The tracking page also includes a description of options
for producing accurate metadata when updating files.) Perhaps others in the community could pursue
(b), either with or without further discussion in this list?

This finesses the heartfelt argument of the subject line: whether ACDD should include those attributes that
can be derived from the data. Successfully pursuing the agreed (?) goals A-C should reduce the intensity of
that issue, and may provide new perspective going forward.

[1] http://wiki.esipfed.org/index.php/NetCDF_Utilities_Metadata_Handling#Table_of_Data_Product_Utilities
```

On 5/27/2014 10:25 PM, John Graybeal wrote:

```
All,

I did a pretty thorough rewrite of the Maintenance of Metadata in Derived Products section, to reflect Steve
Hankin's input (thank you!) and my own further analysis. Below is the new text of the section. While I edited
a lot of Steve's wording, I think the last paragraph advances his provided text pretty well, subject to review of course!

I did not change the actual attribute definitions to reference back to this section, because adding it to some
and not others feels inconsistent, and adding it to more a few feels noisy and detracts from most important message
about attribute maintenance. I suspect this detail may be overdesigning, so if it's contentious let's just go with
whatever majority says we should add the back-reference to (none, temporal, geospatiotemporal, all) at the next meeting.

John

P.S.  The lively discussion has been great on the date_x_modified topic too; I'll try to sum up where we've gotten to
in a day or two, if there are no further comments.

=======

ACDD attributes, like all NetCDF attributes, characterize their containing (parent) granules. As NetCDF data are processed
(e.g., through subsetting or other algorithms), these characteristics can be altered. The software or user processor is
responsible to update these attributes as part of the processing, but some software processes and user practices leave
them unchanged. This affects both consumers and producers of these files, which comprises three roles:

* developers of software tools that process NetCDF files;
* users that create new NetCDF files from existing ones; and
* end users of NetCDF files.
NetCDF file creators (the first two roles) should ensure that the attributes of output files accurately represent those files,
and specifically should not "pass through" any source attribute in unaltered form, unless it is known to remain accurate. NetCDF
 file users (all three roles) should verify critical attribute values, and understand how the source data and metadata were
generated, to be confident the source metadata is current.

The ACDD geospatiotemporal attributes present a special case, as this information is already fully defined by the CF
coordinate variables (the redundant attributes are recommended to simplify access). Errors in these attributes will
create an inconsistency between the metadata and data of the granule or file. The risk of these 'inconsistency errors'
is highest for files that are aggregated into longer or larger products, or subset into shorter or smaller products,
such as files from numerical forecast models and gridded satellite observations. For this reason, some providers of
those data types may choose to omit the ACDD geospatiotemporal attributes from their files. If the ACDD geospatiotemporal
attributes are present, checking them against the CF coordinate variables can serve as a partial test of the metadata's validity.
```

Deprecated Date Fields

Are the date_modified, date_created, date_issued fields deprecated in favor of the following elements date_content_modified, date_values_modified, date_product_generated fields? Are the latter names what we really want? See the Attributes Still Open (http://wiki.esipfed.org/index.php/Talk:Attribute_Convention_for_Data_Discovery_(ACDD)_Working#Attributes_Still_Open) section for details.

...

Summary of Attribute Changes

Attributes Without Comment

Highly Recommended: title, summary

Recommended: id, naming_authority, comment, processing_level, acknowledgment, geospatial_* (bounds, lat_min, lat_max, lon_min, lon_max, vertical_min, vertical_max, vertical_positive), time_coverage_start, time_coverage_end, time_coverage_duration, license (wording reordered) Suggested: geospatial_lat_units, geospatial_lon_units, geospatial_vertical_units, coverage_content_type

Attributes Discussed and Resolved

These attributes should receive extra reviewing attention, as they have most recently changed.

Recommended:

- cdm_data_type: all issues resolved.
- creator, creator_email, publisher, publisher_email: no issue with updates (Nan Galbraith notes: creator might be replaced by a less ambiguous term (OceanSITES is going with principal_investigator as the data 'owner' at this point) What if there is no principal investigator, is this an institution instead of a person?.
- time_coverage_resolution: updated to specify _targeted_ spacing (and preferred format)
- standard_name_vocabulary: someone pointed out this is unnecessary; in CF the standard_name vocabulary is always CF. It's deleted.
- contributor_info: principal objections (ISO 19139) are resolved; while discussion may be needed, but I think satisfactory structural encodings may be found and should be acceptable.
- keywords: chose to leave keywords the wild west, but with some additional options offered; a URI or prefix syntax is also OK, e.g., 'GCMD:space science'
- keywords_vocabulary: in the 'wild west' spirit above, multiple keyword vocabularies can be separated by a comma, and specified in keywords attribute with a prefix (why not?); but we dropped this from Recommended to Suggested
- history: reference to external history metadata can be included; we avoided conflicting with existing one-line-per-process recommendation
- date_content_modified: (was date_modified) the dates explicitly apply to both creation and update
- date_values_modified; new term, like date_content_modified, but applies only to values; the dates apply to both creation and update
- creator_url, publisher url: moved to Suggested, changed to _uri, and specified to apply to person only

Suggested:

- geospatial_*_resolution (lat, lon, vertical): updated to specify _targeted_ spacing
- creator_project, creator_institution, publisher_project, publisher_institution: ok to keep, if Suggested category is downplayed
- creator_project_info, creator_institution_info, publisher_project_info, publisher_institution_info: (deleted ISO 19139) ditto
- date_product_generated: (was date_issued) further updated description to reflect this is when the created file or product was generated (not when a stream first came into existence)
- coverage_content_type: deleted this, it was a recent addition and not strongly needed
- Metadata_Link: defined and made lower case

Deprecated:

- Created this section
- Metadata_Conventions: moved to deprecated, changed text significantly per separate email thread; reference John's email titled Metadata_Conventions and Metadata_Link
- date_created:deleted in favor of date_product_generated (which used to be date_issued); we did not have a use case for knowing the date a stream or product was _first_ generated, once it has been updated
- date_issued: changed name to date_product_generated
- date_modified: changed name to date_content_modified

--Graybeal (talk) 17:12, 19 August 2014 (MST)

Attributes Still Open

- There is still discussion of the replacement of the 'date' terms. Here is the summary from 2014.07.17 email:

```
As a refresher, the original attributes were date_modified, date_created, and date_issued. They were replaced by date_content_modified,
date_values_modified_, and date_product_generated (definitions below).

Two main questions remain unresolved, and a third suggestion.

1) This seems like not enough of a problem, and arguably the solution makes things worse.
2) Should the first item be about ALL content, or just about the NON-VALUES content?
3) The new names date_content_modified and date_values_modified might be clearer as file_values_date and file_date.

Regarding (1): After re-reading all the posts, it seemed a slight majority preferred the change. Many of them were confused by the original names,
and some felt they didn't address all use cases. Those who _didn't_ want the change asserted they had a clear understanding of what the previous
attributes meant.

Regarding (2): The argument for all content was that many users want a date that indicates when the file was last changed in any way.
The argument for non-values content was that it is clearly distinct from the date_values_changed. Not as many letter-writers here, though
the authors preferred it as written.

Regarding 3: Alternative names file_date and file_values_date were proposed. The observation was made that the _values_ attribute needed to clearly
specify whether it included ancillary variables, which are values in their way. While the original choices were made to avoid the use of 'file' in the name
(because not all services are providing files as such), the new names might be clearer anyway.
```

The 3 definitions as they stand after further tweaks by Anna and John G (first 2 are RECOMMENDED, third is SUGGESTED):

date_content_modified

The date on which any of the provided content, including data, metadata, and presented format, was last created or changed (ISO 8601 format)

date_values_modified

The date on which the provided data values were last created or changed; excludes metadata and formatting changes (ISO 8601 format)

date_product_generated

The date on which this data file or product was produced/distributed (ISO 8601 format). While this date is like a file timestamp, the date_content_modified and date_values_modified should be used to assess the age of the contents of the file or product.

--Graybeal (talk) 17:12, 19 August 2014 (MST)

-- Graybeal (talk) 16:44, 3 May 2013 (MDT)

Nan 4/22/2013:

It might be a good idea to cross check against the definitions that NODC has added - as part of their NetCDF template project they wrote some better descriptions. They're at http://www.nodc.noaa.gov/data/formats/netcdf/

There are a few categories of terms that need better definitions, IMHO.

1. people: creator_name (recommended) publisher_name (suggested)

In a 'normal' research/observing/modeling situation, who are these people?

I think there are 2 necessary points of contact, the person who 'owns' the research and gives you the go-ahead to use/publish the data, and the person who put the data into the file and/or on line. You don't really need to know how to contact the other contributors, even if they had equally or more important roles.

I believe that NODC recommends naming the principal investigator as the 'creator' - although in some circumstances there is no single PI, so maybe we should say this is the person who grants the use of the data.

I'm using the publisher as the person who wrote the actual file that contains the terms, and I'm listing co-PIs and data processors as contributors.

Other comments are moved below. jbg

Summary of Changes re Publisher/Creator

I went with Publisher Name, Creator Name, Publisher Info (rich metadata), Creator Info (rich metadata), and Contributor Info (rich metadata); the latter could include owner or any other person/role. All of the 'rich metadata' items could include s role explicitly, presumably from a controlled vocabulary; either the same role or (if you want to create havoc) a different one.

I deleted creator_email and creator_url; if you want to add those, do it in the Info field.

--Graybeal (talk) 19:23, 20 May 2013 (MDT)

Re: *Summary of Changes re Publisher/Creator* -- NanGalbraith (talk) 08:40, 30 July 2013 (MDT)

> I noticed that there was no publisher, just publisher institution etc, so I added publisher with a definition of *The person responsible for the data file, its metadata and format*.

Is that the definition we're using?

I think we have reached consensus that the _info fields are too difficult to parse (Ted's comment); should we go back to _email and _url?

Also, I moved a lot of these out of the 'recommended' category: creator_institution_info, publisher_institution, publisher_institution_info, publisher_project*

One last pitch: with thanks for reminding me, to Mike McCann:

These terms exist in ISO CI_RoleCodes, so why are we not using them?

publisher - The person responsible for the data file, its metadata and format.

principalInvestigator - The person who is responsible for the science content and intellectual property of the dataset

originator - (alternate for principalInvestigator) the person or institution responsible for the science and intellectual property in the dataset, when there is no principalInvestigator

--Graybeal (talk) 14:40, 17 September 2013 (MDT) I'm in favor of using the ISO terms. Note: The definitions above are not the ones I found in ISO; the definitions in ISO are a little bit further down, and are hopelessly self-referencing.

Re: -- Graybeal (talk) 16:48, 3 May 2013 (MDT)

> Ted 4/22/2013: Most of these concerns are discussed at http://wiki.esipfed.org/index.php/NetCDF,_HDF,_and_ISO_Metadata along with more general solutions.

Re: -- Graybeal (talk) 16:48, 3 May 2013 (MDT)

> Nan 4/26/2013: One other item that I think we might need to have - beyond better definitions for some of the existing terms - is a CV for contributor roles. I think one exists, somewhere, but I'm not sure where. BODC, maybe? MMI? Or should this really be free text?

Re: Re: -- Graybeal (talk) 16:49, 3 May 2013 (MDT)

John 4/26/2013: Should be from a controlled vocabulary IMHO. BODC has (for SeaDataNet) an extension of ISO role terms, if I recall correctly. I think it isn't just for contributor roles, it's for all roles that this is needed—ISO wasn't very thorough in the first place, but there will always be new ways for people to be connected to a data set.

I don't think we have to be restrictive (in what roles are allowed) but I think we should try to be explicit (about what a role means).

Re: Re: Re: -- Graybeal (talk) 16:50, 3 May 2013 (MDT)

Ted 4/26/2013: I agree completely that a shared vocabulary with definition is critical. The old ISO vocab is at https://geo-ide.noaa.gov/wiki/index.php?title=ISO_19115_and_19115-2_CodeList_Dictionaries#CI_RoleCode. Many new roles were added in the most recent revision. There is also a brief discussion at http://wiki.esipfed.org/index.php/ISO_People (I will update that list to include revisions)...

What is really important is that the representation allow specification of the source of the code along with the code itself. This is possible in THREDDS, but not ACDD. The job of the standard is to say we use a codelist for this item and that codelist has a location. It is the communities job to say: this is the codelist that our community uses.

Re: Re: Re: Re: -- Graybeal (talk) 16:53, 3 May 2013 (MDT)

Derrick S: Codelists can be seen as antithetical to the CF goal of creating self describing files. Can we figure out a way to encode ISO objects with the need for references to other objects while still staying true to our goal of remaining aligned with CF? The last thing I'd want us to recommend is to open a door down a pathway back to Grib and BUFR.

=Re: Re: Re: Re: Re: -- Graybeal (talk) 17:00, 3 May 2013 (MDT)=

Edward A: Regarding CF, in some ways they already use "codelists", e.g., standard names, so it is not entirely new. Its just their standard names are very human readable at the same time.

=Re: Re: Re: Re: Re: -- Graybeal (talk) 16:58, 3 May 2013 (MDT)=

Nan 4/26/2013: I think we can use terms from a CV, but they should be meaningful, not URLs or those lovely 5 character codes that hark back to languages we've forgotten we ever knew.

We can select one CV, or we can add a term 'rolecode_vocabulary' (that would be fairly reasonable, since we're already using 'keyword_vocabulary').

The SDN roles below are new, but the ISO roles are from a slightly outdated page at NODC. I just find this format easier to look at than the full xml and csv formats that are available on line.

Personally, neither of these is very appealing - I hope the new ISO codes will be better.

SeaDataNet roles

- metadata collator: Responsible for the compilation of metadata for one or more datasets and submission of that metadata to the appropriate SeaDataNet metadata repository.
- programme operation responsibility: Responsible for the operation of a data collecting programme.
- programme archive responsibility: Responsible for the archive centre handling distribution of delayed mode data from a collecting programme and the long term stewardship of its data.
- programme realtime responsibility: Responsible for the centre handling distribution of true and near real time data from a collecting programme.
- contact point: Person responsible for the provision of information in response to queries concerning the metadata or underlying data.
- principal funder: Person or organisation that funds the majority of an activity. contributing funder: Person or organisation that contributes to the funding of an activity.
- principal investigator: Scientific lead of data collection within a programme

ISO roles

- resourceProvider: party that supplies the resource
- custodian: party that accepts accountability and responsability for the data and ensures appropriate care and maintenance of the resource
- owner: party that owns the resource
- sponsor: party that sponsors the resource
- user: party who uses the resource
- distributor: party who distributes the resource
- originator: party who created the resource
- pointOfContact: party who can be contacted for acquiring knowledge about or acquisition of the resource
- principalInvestigator: key party responsible for gathering information and conducting research
- processor: party who has processed the data in a manner such that the resource has been modified
- publisher: party who published the resource
- author: party who authored the resource
- collaborator: party who conducted or contributed to the research

==Re: Re: Re: Re: Re: Re: -- Graybeal (talk) 18:03, 3 May 2013 (MDT)==

Thanks for aggregating these terms. I agree none of these role vocabularies are very appealing, but I suspect that's because the world they describe is messy. I do not see how a single vocabulary can satisfy everyone's needs, especially for keywords; nor how naming the vocabulary title creates an unambiguous reference that everyone can use to look up terms from it. I guess I'm just stuck on the lack of provided functionality in this respect.

==Re: Re: Re: Re: Re: Re: -- Graybeal (talk) 17:10, 3 May 2013 (MDT)==

Ted H 4/27/2013: The suggestion to add an attribute called rolecode_vocabulary demonstrates very well the problem with this approach

- a community has a documentation need and, in order to address that need, we need to add a new concept into the convention. Do we end up with a *_vocabulary attribute for every attribute that can benefit from a shared vocabulary? I think this would be difficult to maintain.

As an alternative, we create a responsibleParty type group that includes a role from a shared vocabulary and information that describes people or organizations. The role has a value and a source which is the shared vocabulary that it comes from.

Are we a community of convention users or convention developers? When we say we need a mechanism for describing responsibleParties that includes a role from a shared vocabulary and descriptive information, we are convention developers. When we say we need a vocabulary to describe roles like principleInvenstigator or instrumentDeveloper, we are acting as a community using a convention.

What I am trying to do is separate these two roles so that when a community says "we need a shared vocabulary for x", we do not have to add a new attribute called x_vocabulary to the convention.

Re: -- Graybeal (talk) 17:09, 3 May 2013 (MDT)

Ken C 4/27/2013: All we say at NODC in our netCDF templates for the creator_ attributes is copied below… we discussed attributes like this a lot when documenting our templates and finally "settled" on the idea of creator being associated with "collector" of the data. Of course even that is not perfect. We don't say anything about PIs, since as Nan points out there is often no single PI. I would add that there is often no PI at all… many, many, datasets come to us now as a result of sustained and operational observing programs and systems, where the idea of a "PI" itself doesn't even apply.

* creator_email: Email address of the person or institution that collected the data. -- The email of the person or institution may be found in the NODC tables for persons (http://www.nodc.noaa.gov/cgi-bin/OAS/prd/person) and institutions(http://www.nodc.noaa.gov/cgi-bin/OAS/prd/institution). Use the short name of the institution if available.
* creator_name: Name of the person who collected the data. -- Use the name from the NODC persons(http://www.nodc.noaa.gov/cgi-bin/OAS/prd/person) table when applicable.
* creator_url: The URL of the institution that collected the data. -- The url of the institution can be found in the NODC institutions (http://www.nodc.noaa.gov/cgi-bin/OAS/prd/institution) table

-- Graybeal (talk) 16:44, 3 May 2013 (MDT)

Nan 4/22/2013: There are a few categories of terms that need better definitions, IMHO. *(continued)*

2. file times

- date_created (recommended)
- date_modified (suggested)
- date_issued (suggested)

These could well have different meanings for model data; for my in situ data, I have 2 (or, for real time data, possibly 3) useful file times; the time the last edit or processing occurred, which is the version information and could be useful if the underlying data has been changed, and the time the file was written, which could provide information about translation errors being corrected. (We don't update files, we overwrite them; some people might need to describe the time the original file was written and time of last update?) For real time data it could also be interesting to know the last time new data arrived, which could be asynchronous.

NODC doesn't seem to use date_issued, but they have defs for created and modified.

- date_created: "The date or date and time when the file was created.... This time stamp will never change, even when modifying the file."
- date_modified: This time stamp will change any time information is changed in this file.

Summary of Changes re File Times

If there is the concept of date_modified, it has to be the last time the data changed (as the public sees it). That's the most important metadata of all, so now it's in the Recommended section.

If that is date_modified, then date_created has to be the original creation date, when information was first available on this file.

I could not think of a non-bizarre use case for date_issued, so I deleted it.

--Graybeal (talk) 19:25, 20 May 2013 (MDT)

3. Keywords

Since iso uses keyword type codes instead of cramming all the possible keywords (theme, place, etc) into one structure, I don't see why we don't do something similar. We could use our pseudo-groups syntax; keywords_theme, keywords_dataCenter ...etc.

Summary of Changes re Keywords

I created an arcane way to specify multiple keyword vocabularies, and implicitly allowed it to specify prefixes for the keyword field (e.g., "CF:air_temperature, IOOS_Key:Nutrients, My Favorite Keyword, AirTemperature"). I opened up the format (it's free text, why not), which leaves the battle to be fought over best practices.

--Graybeal (talk) 19:30, 20 May 2013 (MDT)

Re: 3. Keywords -- Graybeal (talk) 18:13, 3 May 2013 (MDT)

Not sure how the type codes are being considered in this context, as additional attributes or as an organizing technique inside the keywords attribute?

I consider it a fail that there is no agreed way to support two keyword vocabularies. I therefore propose the following: If a keyword is a URI, it does not have to be a member of the Keyword Vocabulary (because its vocabulary can be derived through other means).

> I wish there were a way that Keywords and Keyword Vocabulary could have a default treatment that makes these two fields fully computer-friendly. Could we permit the Keyword Vocabulary format to be a URI, or to be specified as Name|URI, wiki-like.

4. coordinate 'resolution' terms

The word resolution is a poor choice, and if it's going to be kept, it needs to be defined as meaning 'spacing' or 'shape' and not an indication of the precision of the coordinate. For measurements that are irregularly spaced along a mooring line, it's fairly useless - unless we come up with a vocabulary describing this and other possible values.

For my data, the term might be more useful with the other definition; our depths are approximate 'target depths', and, while we may know the lat/long of an anchor and of a buoy (the latter being a time series, the former being a single point) we don't actually know the lat/long of any given instrument on a mooring line. The watch circle of the buoy is really the 'resolution' we need to supply here.

Re: 4. coordinate 'resolution' terms -- Graybeal (talk) 18:27, 3 May 2013 (MDT)

> Ooh, good point. I think in context of geospatiotemporal *coverage*, 'resolution' is a meaningful word, but without a definition it's wide open to misinterpretation.
>
> Your need is in regard to the measurements/locations provided for the data, right? The three terms that often get used to satisfy your need are precision, accuracy, and error. Can they be specified by the corresponding variable attributes?

-- Graybeal (talk) 18:31, 3 May 2013 (MDT)

Adding Guidance

Do we want to provide any guidance, in addition to the definition?

Re: Adding Guidance -- Ted.Habermann (talk) 09:36, 5 May 2013 (MDT)

> Guidance is incredibly important on many levels. I think it is really important to integrate the guidance into the conformance tool. We have done this more in the ISO rubric then in the ACDD rubric. The rubric results include the links to the guidance and examples... This ends up providing an integrated evaluate / improve environment...

Computability

I often try to make the definition of a parameter clear enough that a computer could recognize and do something with the answer. Is that strongly desirable, weakly desirable, or not of interest?

Summary of Approach re Computability

Some of us find it strongly desirable, but not enough to enforce it throughout. So I added it as an option in a number of places, and tried to encourage it with some of the definitions.

--Graybeal (talk) 19:32, 20 May 2013 (MDT)

Cross-Referencing

There are other pages with guidance and discussion about these terms. Do we want to refer the user explicitly to them, either in the document as a whole or in specific terms?

Re: Cross-Referencing -- Ted.Habermann (talk) 09:37, 5 May 2013 (MDT)

> See Guidance discussion above

Roles-by-Position vs. Roles-by-Code -- Ted.Habermann (talk) 09:16, 5 May 2013 (MDT)

Organizations and people play many roles in the scientific data life-cycle. There are two ways that those roles can be reflected in a metadata record: by position and by code. Many metadata managers are familiar with the roles by position approach because it is used in the FGDC CSDGM. The person referenced from the metadata section is the metadata contact, the person referenced from the distribution section is the distributor, and so on. Using this approach means that the object that holds information about organizations/people does not need a role indicator. That information is inferred by the position in the structure.

The ISO Standards combine the roles-by-position approach with the roles-by-code approach. Roles can generally be inferred from the positions of CI_ResponsibleParty objects in the structure, but flexibility is increased by adding a code for role to the each object. This is helpful when citing a dataset that involves people in multiple roles (principle investigator, publisher, author, resourceProvider) or when specifying the point of contact for a particular section.

The roles-by-position approach allows the roles of the people involved with a dataset to be known when they are accessed separately. For example, the xPath /gmi:MI_Metadata/gmd:contact can be used if one were interested in the metadata contact for a resource. A more general xPath (//gmd:CI_ResponsibleParty) can be used to answer the question “what people or organizations are associated with this dataset”. In the latter case, the role code provides information about roles even though the people are being accessed independent of the structure.

Multiple CI_ResponsibleParties can be included in almost all ISO objects that can include CI_ResponsibleParties. In those cases, roleCodes can be used to associate appropriate roles with particular organizations people if necessary. For example, the ISO CI_Citation object is used to refer to a variety of resources that are not included in a metadata record. It is modeled after a bibliographic reference and can include any number of organizations or people (CI_ResponsibleParties) in any roles. Typically a CI_Citation includes originators or authors and a publisher.

Re: Roles-by-Position vs. Roles-by-Code -- Ted.Habermann (talk) 09:45, 5 May 2013 (MDT)

> The discussion of role codes is interesting from many points of view. The lack of groups in the netCDF model essentially eliminates both of these approaches from

> consideration. There is no structure to attach organizations or people to and there are no objects to attach roles to. The only remaining alternative is the "named element" approach in which the name of the element includes the role. Are there advantages to that?

creator_name and institution definitons. -- Dpsnowden (talk) 13:05, 9 May 2013 (MDT)

The definition of creator_name is now

creator_name</dt>

> The data creator's name, URL, and email. The "institution" attribute will be used if the "creator_name" attribute does not exist.</dd>

The discussion about the roles for individuals is elsewhere in the document. My point here is that the second sentence of the existing definition includes a description of some action that will be taken. While many of us know that the actor in this case is ncISO, not everyone does. Further, we're conflating two concepts, the definition of a term and the use of that term in a particular use case (i.e. translation to ISO 10115* via ncISO). I propose that for this definition in particular and for the entire wiki in general, that we strive to separate these two concepts in the text. Let's first state what ACDD is, and what each term means, and then state one of the admittedly most common use cases.

Summary of Approach re Using Terms in Use Cases

Strove to separate the concept of how it is used from the concept of a term's definition. (One place you can't do that is in the cdm_feature term, which is very explicit about its connection to THREDDS features.)

--Graybeal (talk) 19:34, 20 May 2013 (MDT)

Feature Types (cdm and otherwise) -- Graybeal (talk) 17:40, 20 May 2013 (MDT)

The Unidata ACDD says

> The "cdm_data_type" attribute gives the THREDDS data type appropriate for this dataset. E.g., "Grid", "Image", "Station", "Trajectory", "Radial". Its use is recommended.

The NOAA ACDD says

> The THREDDS data type appropriate for this dataset

This is what ours currently says.

The NODC guidance says

> This attribute is used by THREDDS to identify the feature type, what THREDDS calls a "dataType". The current choices are: Grid, Image, Station, Swath, and Trajectory. These data types do not map equally to the CF feature types. If the CF feature type = Trajectory Time Series, use "Trajectory"; if Point, Profile, or Time Series Profile, use "Station".

The actual THREDDS list is called either dataTypes (code) or dataType Types (doc header), and has the same 5 types listed in the NODC guidance.

If you look up "netcdf feature type" the first link is http://www.unidata.ucar.edu/software/netcdf-java/reference/FeatureDatasets/Overview.html, which says the choices are ANY, NONE; GRID, RADIAL, SWATH, IMAGE; and ANY_POINT, which encompasses POINT, PROFILE, SECTION, STATION, STATION_PROFILE, and TRAJECTORY.

I went with something NODC-like, though it killed me not to include radial, station_profile, etc.

Re: Feature Types (cdm and otherwise) -- NanGalbraith (talk) 13:15, 9 September 2013 (MDT)

featureType is a special NetCDF attribute in CF; it gives the type of Discrete Sampling Geometry, and its presence indicates that the file contains DSG features. This opens a whole set of expectations for the file contents, and some limitations on the dimensions and coordinates allowed. We should stick with cdm_data_type, in my opinion - although I have to ask if it is actually a discovery attribute.

Re: Re: Feature Types (cdm and otherwise) -- NanGalbraith (talk) 08:44, 30 September 2013 (MDT)

> > The term cdm_datatype seems to have originated with ACDD, and it's a poor choice of terms, IMHO, since most THREDDS docs use 'data type' to mean float/int etc. Also, we might want to point to the actual unidata document that defines what we are calling cdm_data_types, at http://www.unidata.ucar.edu/software/thredds/current/netcdf-java/tutorial/PointDatatype.html That page uses the term Observation Datatypes, which is not really any more explicit than cdm_data_type. Feature type is more descriptive, but (as above) it's an overloaded CF attribute.

From the unidata page linked above, these are the definitions of the types:

"Several types of observation collections are described in the Common Data Model's Scientific Datatype layer. A Point Observation dataset contains observations which are not necessarily related in space or time. A Station Observation dataset contains time series of observations at named locations called stations. A trajectory is a collection of observations which are connected along a one dimensional track in space, with time increasing monotonically along the track. A Trajectory Observation dataset contains one or more trajectories."

Depth (!) -- Graybeal (talk) 19:17, 20 May 2013 (MDT)

Depth is fraught.

(0) Vertical positive: I almost made this required. Instead, I moved it from Suggested to Recommended. Obvious reasons.

(1) Vertical min/max: I didn't see in casual inspection a clear practice for min/max specification as a function of vertical_direction_positive = up or down. So I reused a convention established, after long thought, by OOI CI, and documented here (https://confluence.oceanobservatories.org/display/CIDev/Coordinate+Systems+and+Coordinate+Transformations#CoordinateSystemsandCoordinateTransformations-

Vertical) . Trust me, there is one other option for a convention, and it is at least as confusing if not more so.

(2) Vertical units: I assume we are not going to insist on depth as the only vertical coordinate, so I explicitly mention pressure and the use of bar.

People and Institutions -- Ted.Habermann (talk) 13:55, 4 June 2013 (MDT)

The definitions that John proposed are helpful, but raise several issues. Before, we had eight attributes with roles embedded in their names (creator_name, _url, _email, publisher_name, _url, _email, contributor_name, _role) now we have twelve proposed. Many of these proposals would encourage the concatenation of multiple information elements into single fields (contributor_info, ...) with a recommendation of using vcard, ISO 19139 or free text. I am not aware of a mechanism for including ISO 19139 in netCDF attributes. Remember that NcML has the content as XML attributes which makes it fundamentally impossible to embed XML in them and very ugly to embed delimited text. This makes it likely that freetext would be the format of choice. This creates information blobs that are many times difficult to untangle and use, particularly for machines. It is also not clear how we deal with datasets that have multiple creators from multiple institutions. This is a very common circumstance these days. I am not aware of a mechanism for connecting appropriate creator_persons to appropriate creator_institutions when there are multiple occurrences of each. In fact, I do not know of an unambiguous way to include multiple creators in netCDF as it is currently implemented.

Re: People and Institutions -- NanGalbraith (talk) 13:09, 6 September 2013 (MDT)

I replaced _info fields with _url and _email for creator and publisher, because I agree that these are easier to parse. I would like to move the _url fields (along with a few others) from the Recommended section to Suggested, or possibly to add a category that isn't so much suggested as ... *might be to be considered*. The creator_institution_info, creator_project*, publisher_institution*, and publisher_project* fields don't aid in discovery enough to include them, in my opinion.

Conventions or Metadata_Convention -- NanGalbraith (talk) 09:40, 19 November 2013 (MST)

We need to discuss whether to remove the existing Metadata_Conventions attribute and add ACDD-1.3 (or other) to the 'Conventions' attribute, as is recommended by the unidata guidance.

From Writing NetCDF Files: Best Practices and other unidata guidance documents:

If present, Conventions is a global attribute that is a character array for the name of the conventions followed by the dataset.

The `Conventions' attribute may be a single text string containing a list of the convention names separated by blank space (recommended) or commas (if a convention name contains blanks)

Document the convention you are using by adding the global attribute "Conventions" to each netCDF file, for example:

> Conventions = "CF-1.3";

This is under discussion on the ACDD team email:

'I have always preferred the idea of using the "Conventions" attribute rather than "Metadata_Conventions". However, client support for multiple values in the "Conventions" attribute was not very good back when ACDD was originally written. And, while explicit mention of multiple values in the "Conventions" attribute have been in the NUG for some time, it is (I believe) only now slated for the next version of CF [1].

Does anyone have a good sense of client support for this now?

Then again, there's the chicken and egg issue. Clients will be slow to support this feature until someone starts producing data that uses this feature.' - Ethan

'We should discuss the deprecation of Metadata_Conventions more closely at the next telcon. We for one are using it currently in many, many GHRSST granules.' - Ed Armstrong

Retrieved from "http://wiki.esipfed.org/index.php?title=Talk:Attribute_Convention_for_Data_Discovery_1-2_Working&oldid=47245"

- This page was last modified on 19 August 2014, at 23:32.
- This page has been accessed 1,486 times.
-