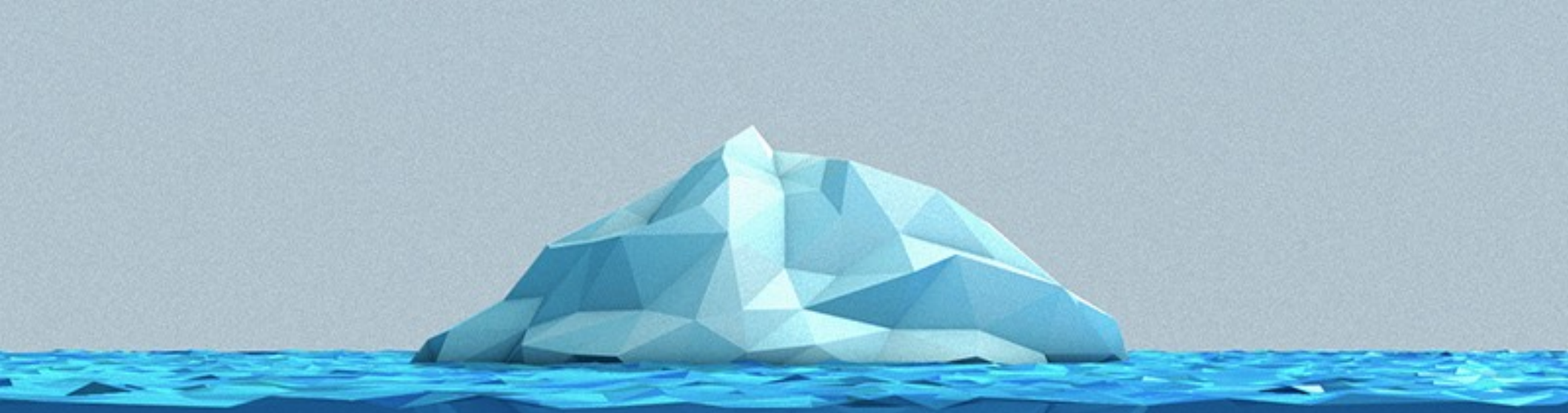


Data structure/formatting

Øystein Godøy and Markus Fiebig

NorDataNet
Norwegian Scientific Data Network

Outline

- Standard names, vocabularies
- NetCDF/CF grid, trajectory, profile, timeseries
- Granularity requirements

NorDataNet
Norwegian Scientific Data Network

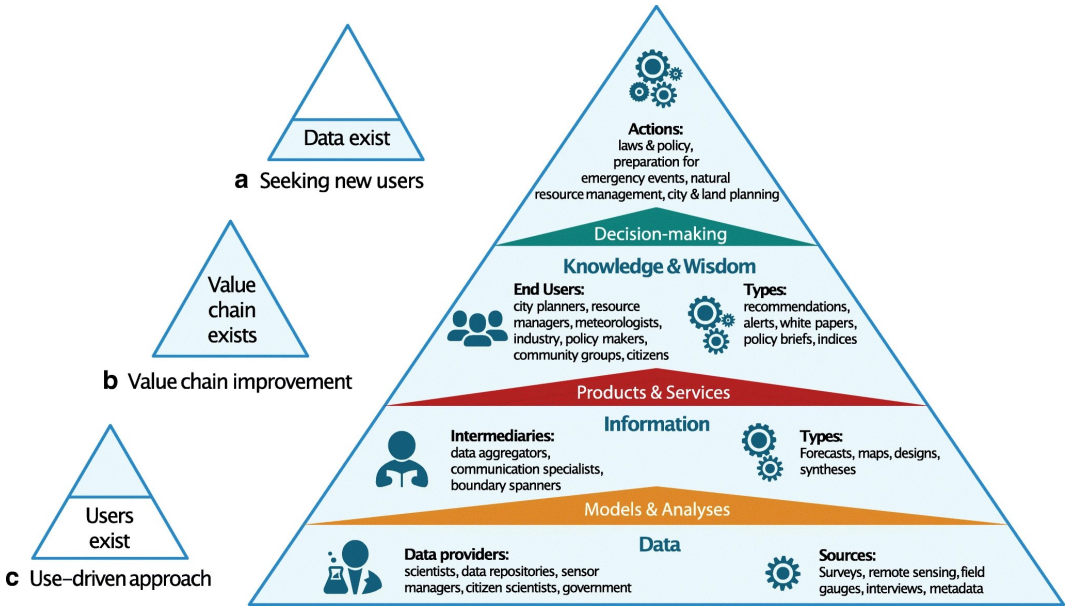
Benefits of standardised documentation

- Why not use the “Google” approach?
- Science is based on a shared terminology
  - There will never be only one proper way of documenting
  - There will always be a need for brokering
- Data and metadata must be connected
  - To find data
  - To use data
- Standardised documentation and formatting
  - enables the possibility to filter datasets
  - enables the possibility to link datasets
  - enables standardised applications to analyse data
  - enables users to use the data
- Need to be pragmatic…
  - And let computers do the boring part
  - But humans need to instruct computers

NorDataNet
Norwegian Scientific Data Network

Standardised documentation criteria

- Self contained information
  - Structural representation of data
  - Semantic annotation of data and structures
  - Avoid using “containers”
  - Slicing of data during analysis
  - Sustainability of format specification, tools and APIs



Virapongse, A., Pearlman, F., Pearlman, J. et al. Ten rules to increase the societal value of earth observations. Earth Sci Inform 13, 233–247 (2020). https://doi.org/10.1007/s12145-020-00453-w

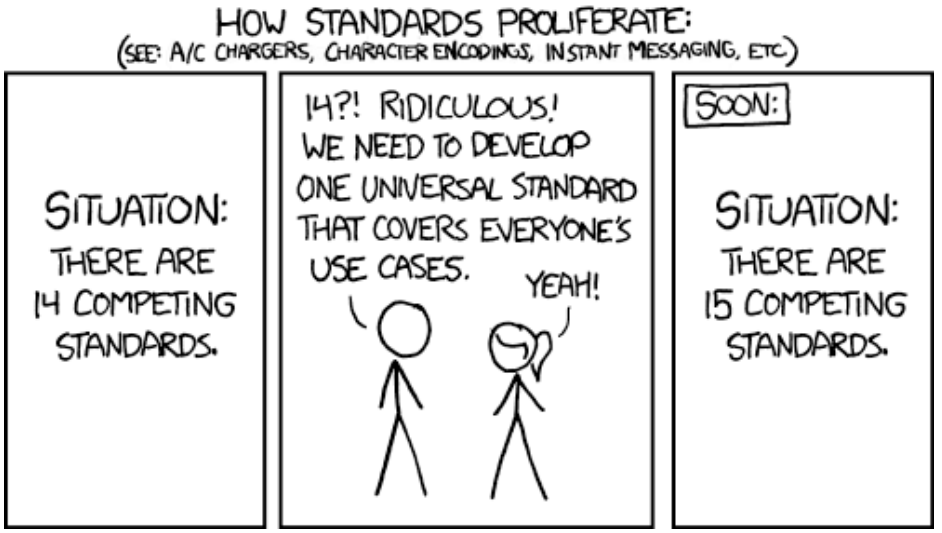
NorDataNet
Norwegian Scientific Data Network

Data Formats: Choosing and Adopting Community Accepted Standards

- Most projects (rightly so) focus on the content of their data files, you need to consider the format as well.
- Since you captured or created the data, and stored them in your own files, you know
  - how the data are organized,
  - how to read them,
  - how to use them,
  - characteristics of the data that could constrain their use.
- **The goal of a good data format is to make it easier for others to read the data too.**
- Many hours have gone into developing standards for formats – try to learn from them.

Why use community standards

- If you try to develop your data format from scratch, you will forget something.
- Build on the experience and improvements built into the community standards over years of use.
- Tools and analysis software natively support reading community standard data.
- Reduce development effort and support reuse.
- Positive feedback – they are more likely to be adopted by others.



http://xkcd.com/927/

Curt Tilmes, NASA, 2013 ESIP Material

There is nothing like a perfect standard...

NorDataNet
Norwegian Scientific Data Network

Use self describing data formats

- Self-describing data formats have become a well accepted way of archiving and disseminating scientific data.
- Before self-describing data formats became widely used, each project often invented their own data formats, often raw binary or even ASCII.
- These approaches had a number of problems:
  - Machine dependent byte ordering or floating point organizations
  - Required a ‘key’ to be able to open the file and read the right data.
  - A new custom reader is needed for each different data organization. Working in a new language could be very difficult since you have to redevelop the reader anew.

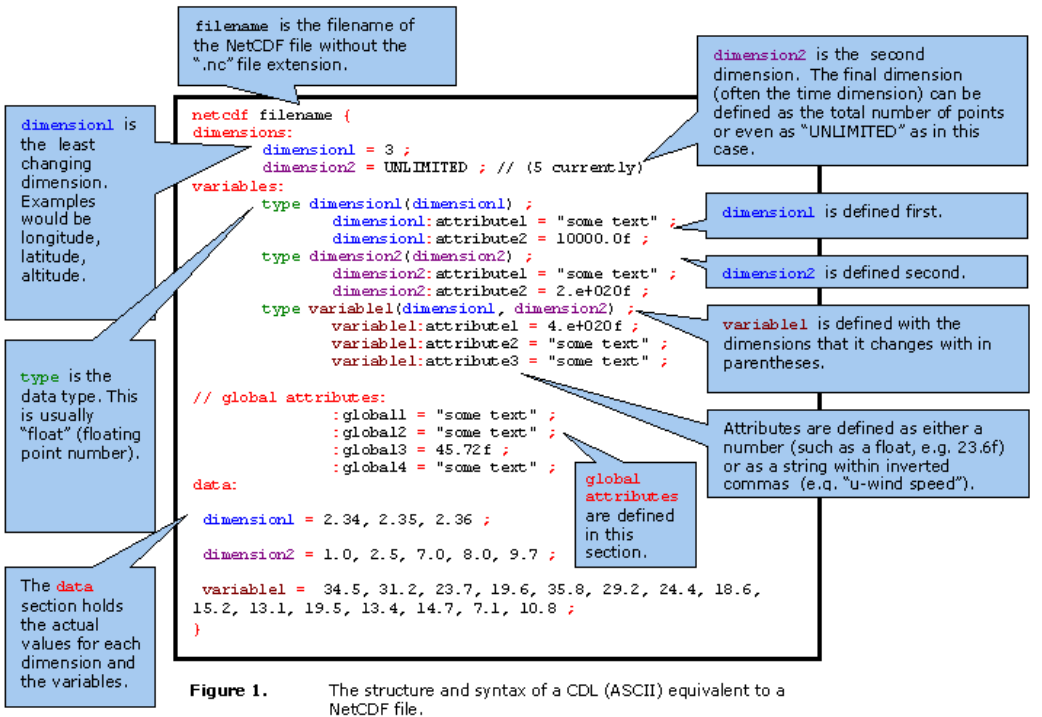
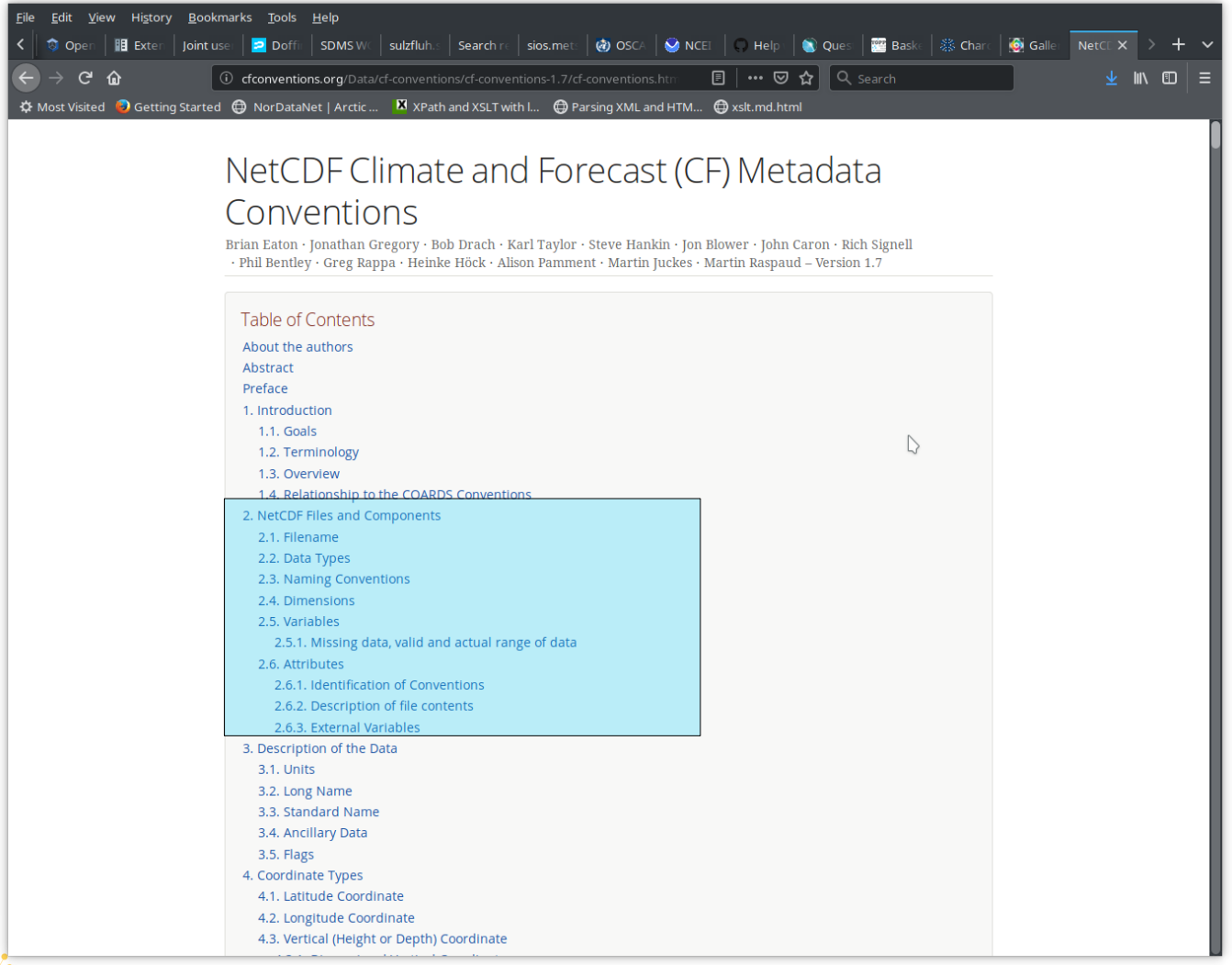
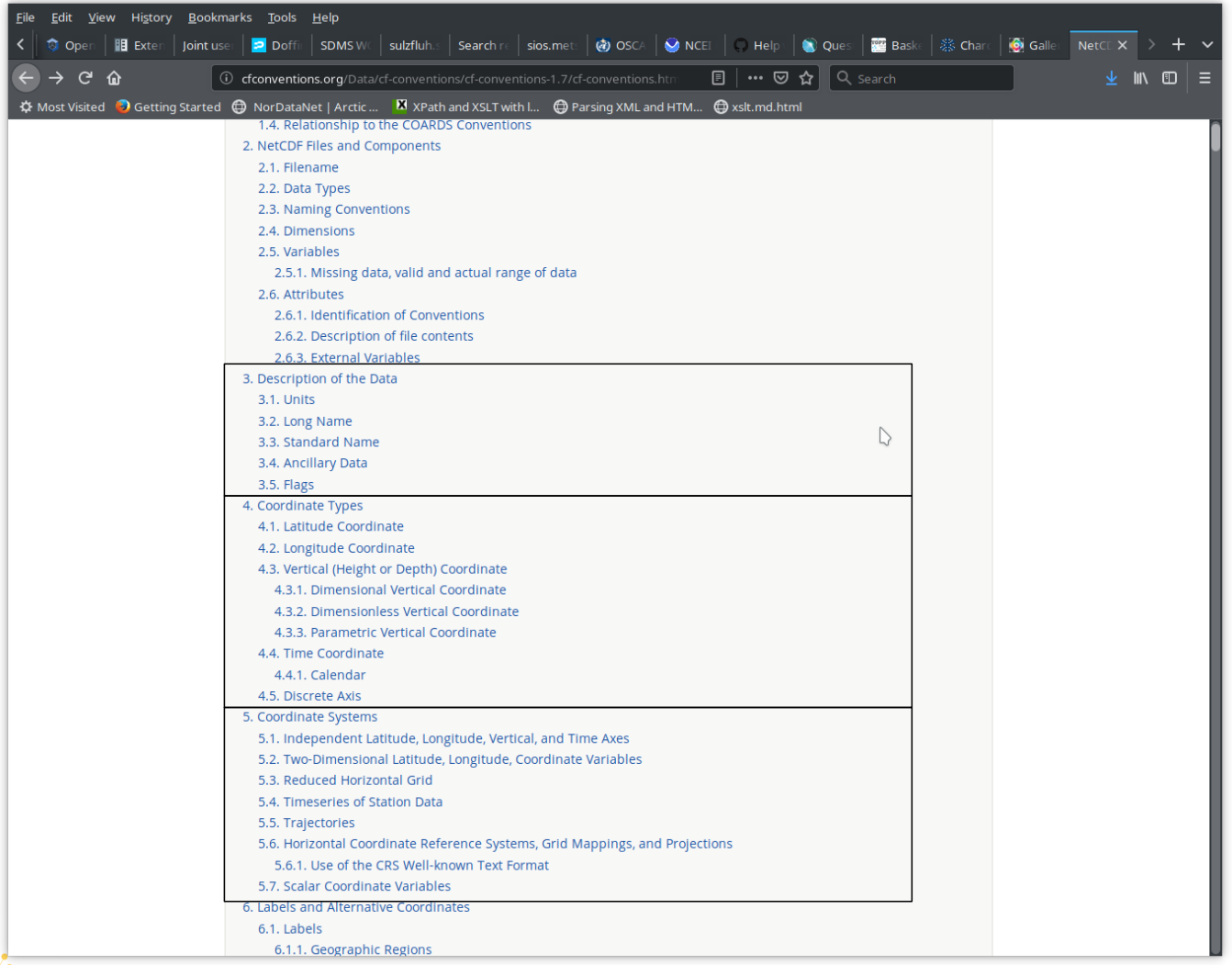


Figure 1. The structure and syntax of a CDL (ASCII) equivalent to a NetCDF file.

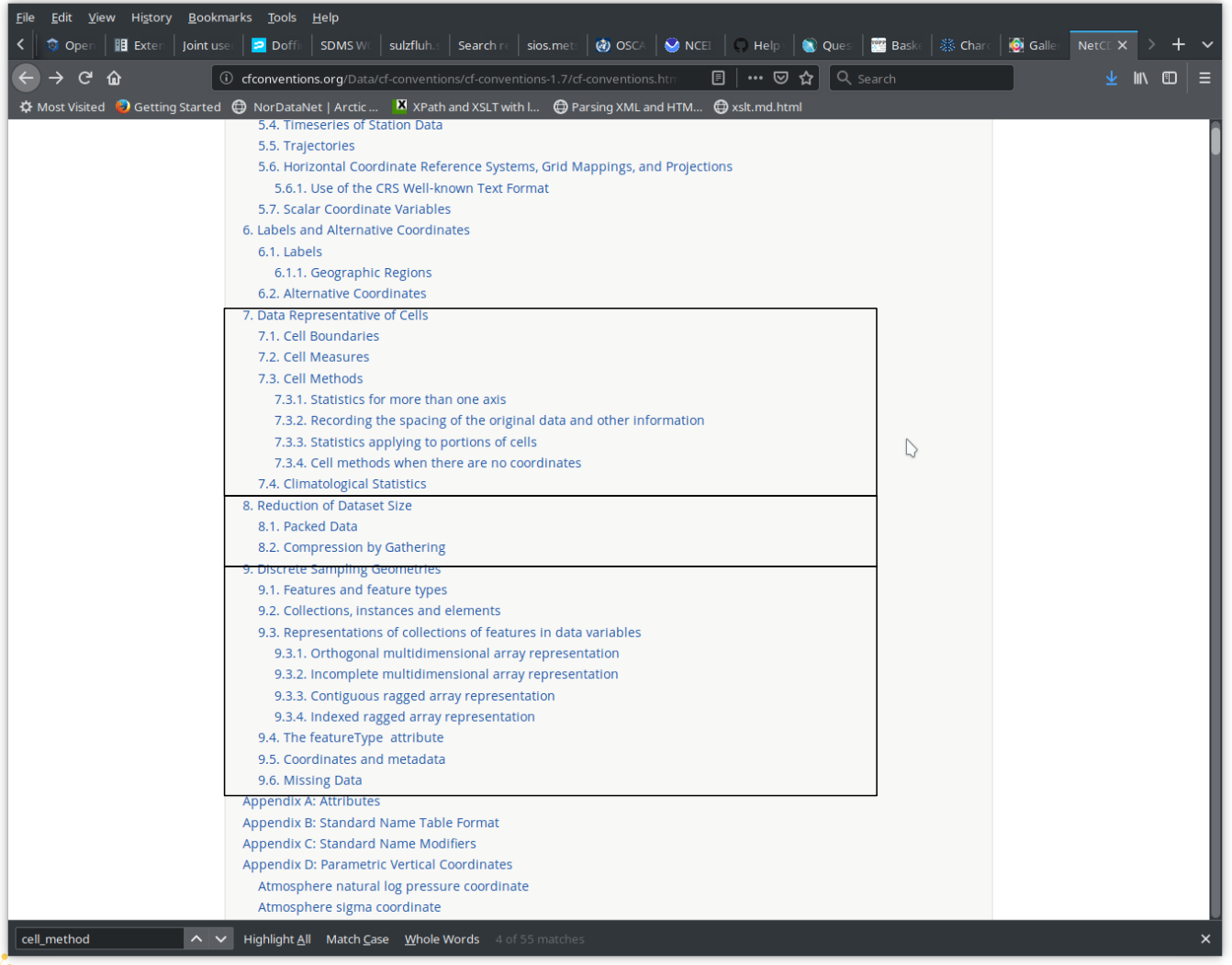
Curt Tilmes, NASA, 2013 ESIP Material



NorDataNet
Norwegian Scientific Data Network

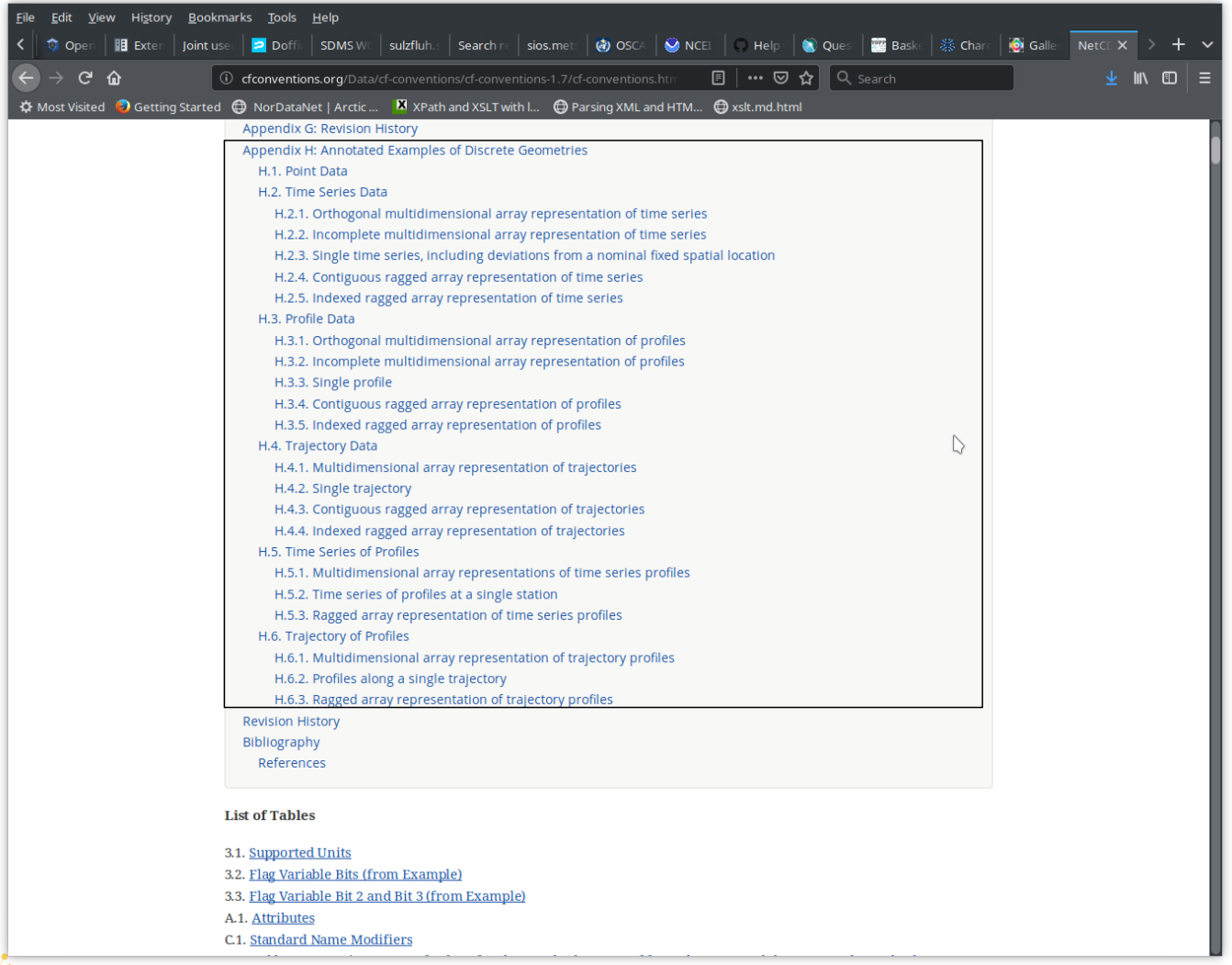


NorDataNet
Norwegian Scientific Data Network

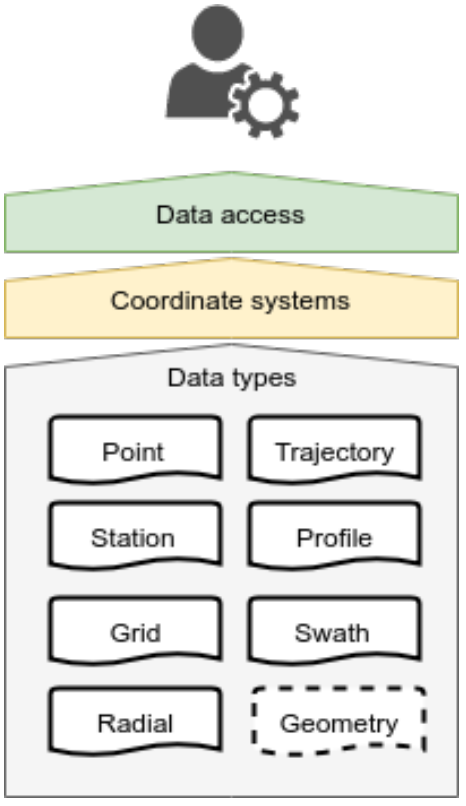


NorDataNet
Norwegian Scientific Data Network

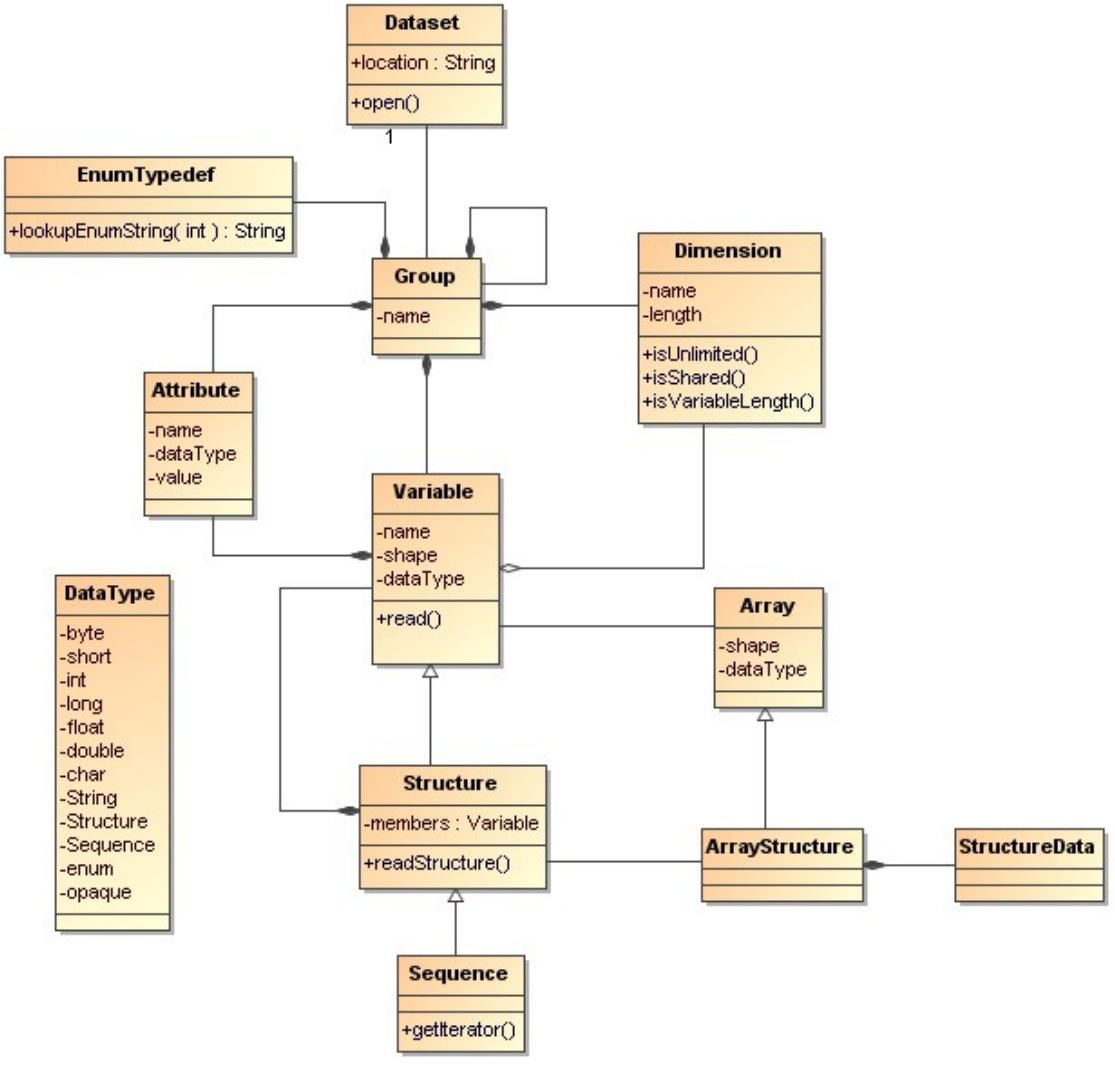
featureType	Description of a single feature with this discrete sampling geometry		Link
	Form of a data variable containing values defined on a collection of these features	**Mandatory space-time coordinates for a collection of these features**	
point	a single data point (having no implied coordinate relationship to other points)		
	data(i)	x(i) y(i) t(i)	Section H.1, "Point Data"
timeSeries	a series of data points at the same spatial location with monotonically increasing times		
	data(i,o)	x(i) y(i) t(i,o)	Section H.2, "Time Series Data"
trajectory	a series of data points along a path through space with monotonically increasing times		
	data(i,o)	x(i,o) y(i,o) t(i,o)	Section H.4, "Trajectory Data"
profile	an ordered set of data points along a vertical line at a fixed horizontal position and fixed time		
	data(i,o)	x(i) y(i) z(i,o) t(i)	Section H.3, "Profile Data"
timeSeriesProfile	a series of profile features at the same horizontal position with monotonically increasing times		
	data(i,p,o)	x(i) y(i) z(i,p,o) t(i,p)	Section H.5, "Time Series of Profiles"
trajectoryProfile	a series of profile features located at points ordered along a trajectory		
	data(i,p,o)	x(i,p) y(i,p) z(i,p,o) t(i,p)	Section H.6, "Trajectory of Profiles"



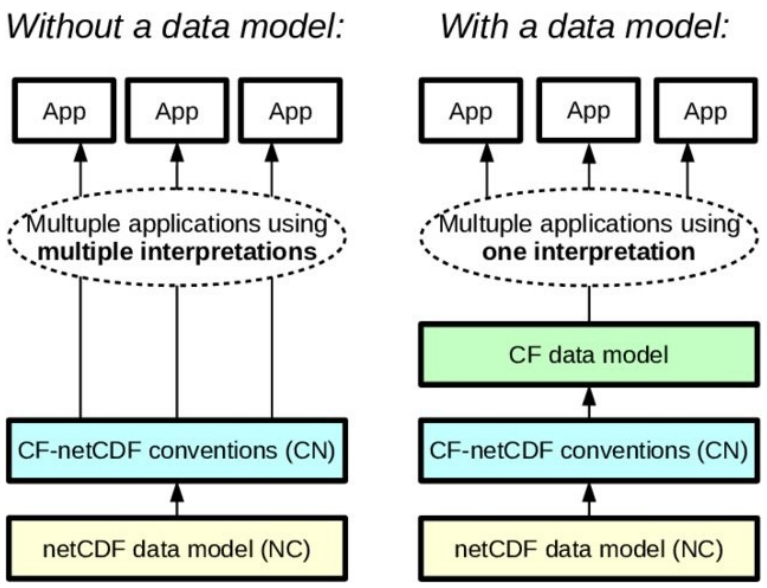
NorDataNet
Norwegian Scientific Data Network

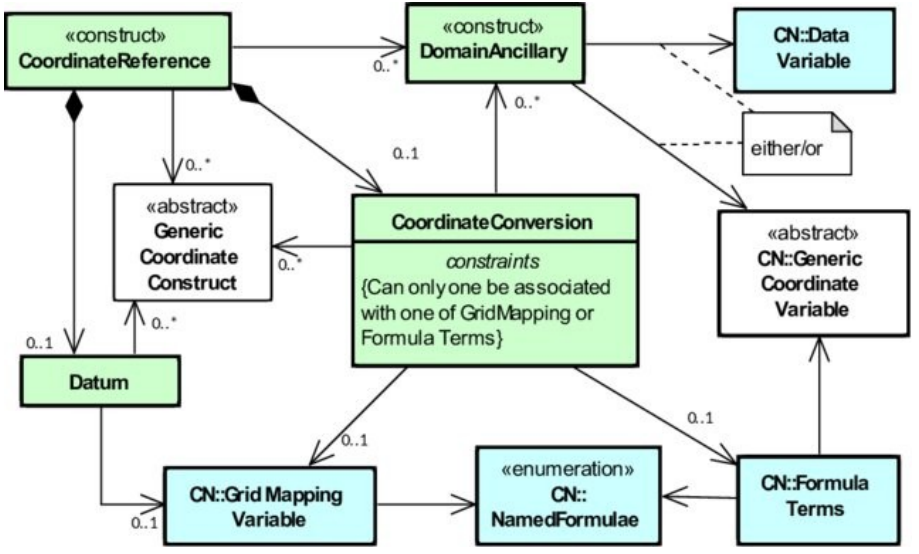


NorDataNet
Norwegian Scientific Data Network

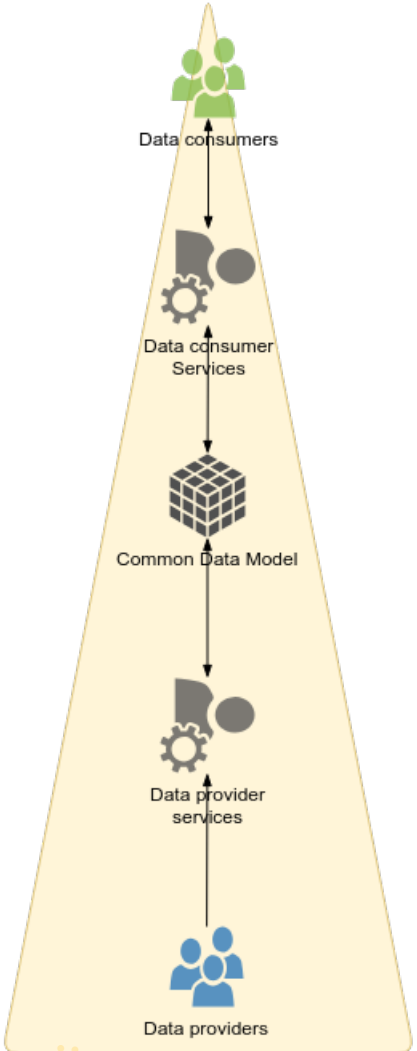


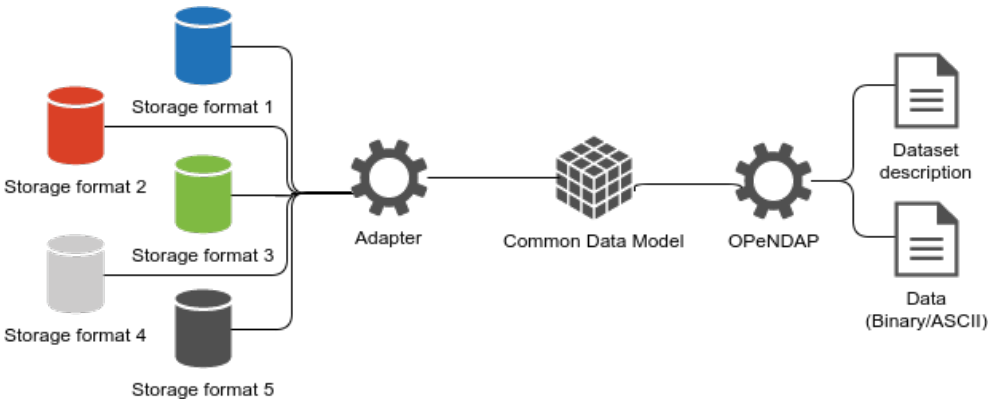
NorDataNet
Norwegian Scientific Data Network



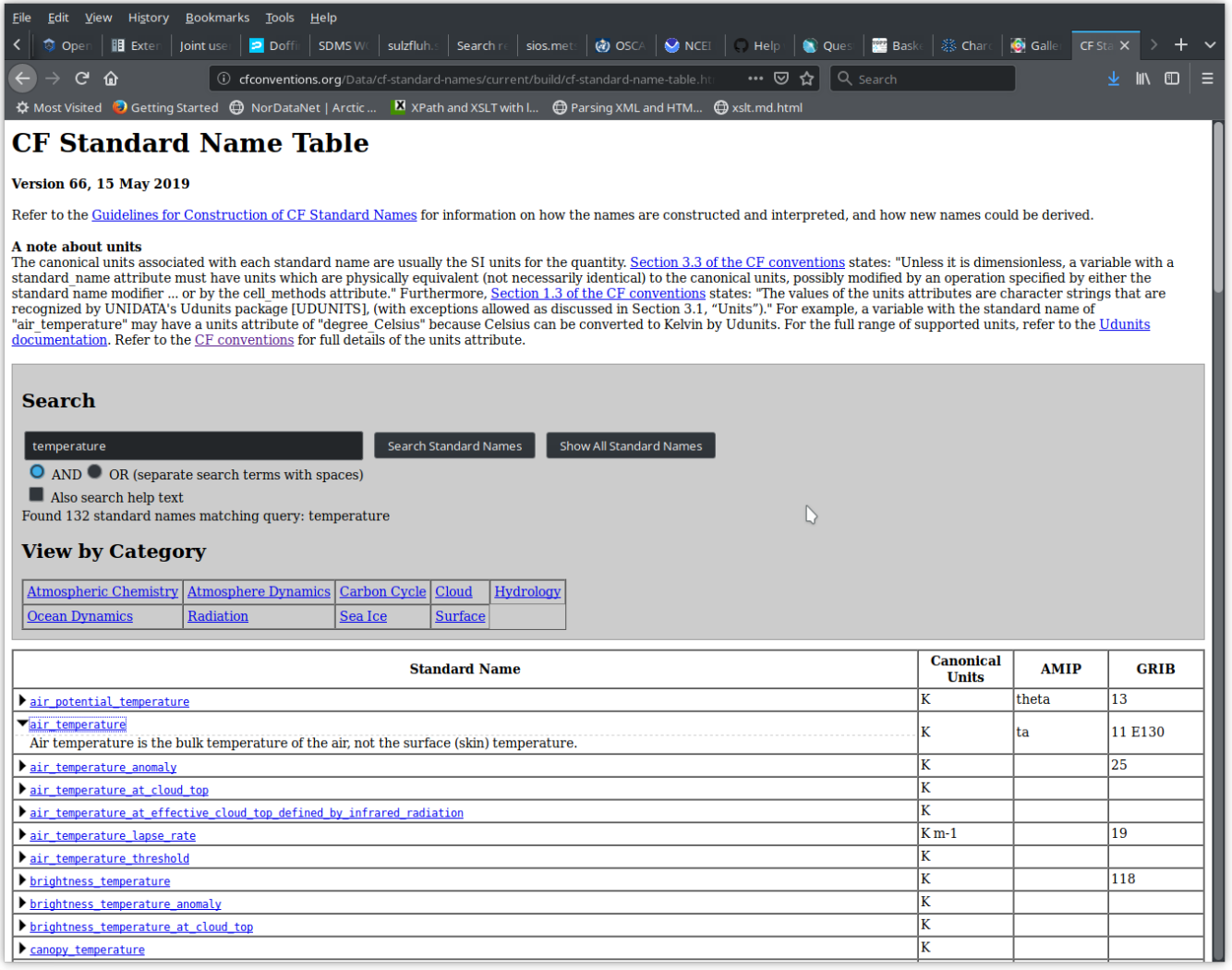


https://www.researchgate.net/publication/318324043_A_CF_data_model_and_implementation/figures?lo=1

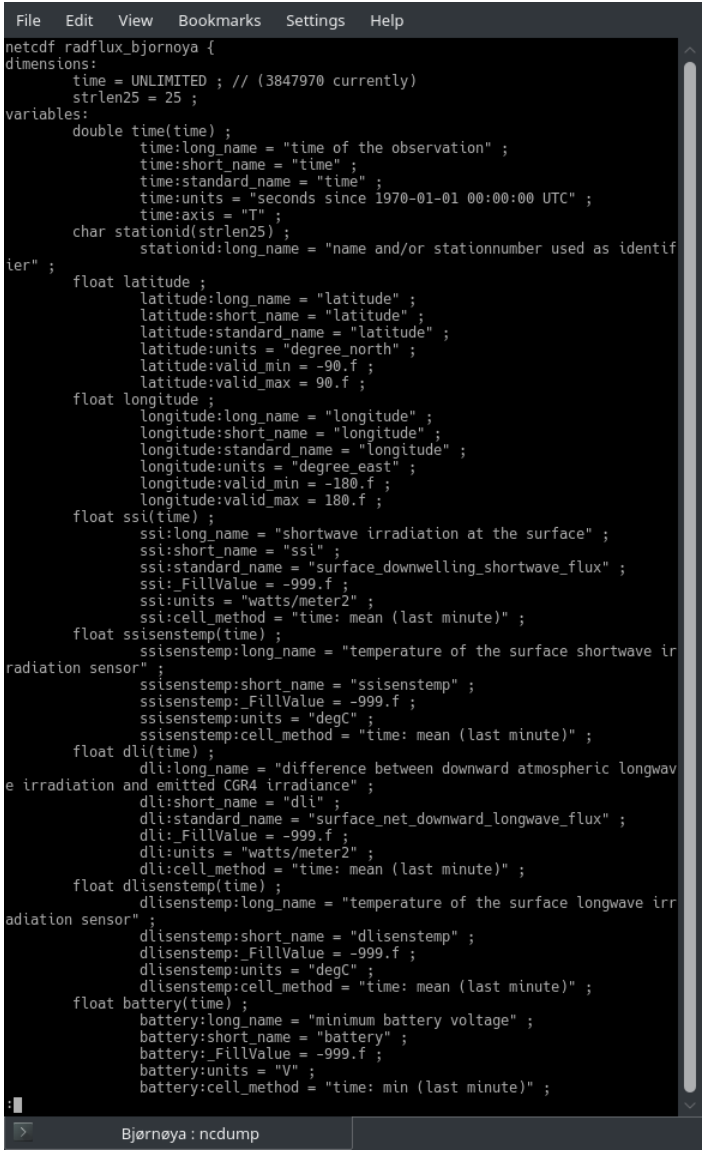


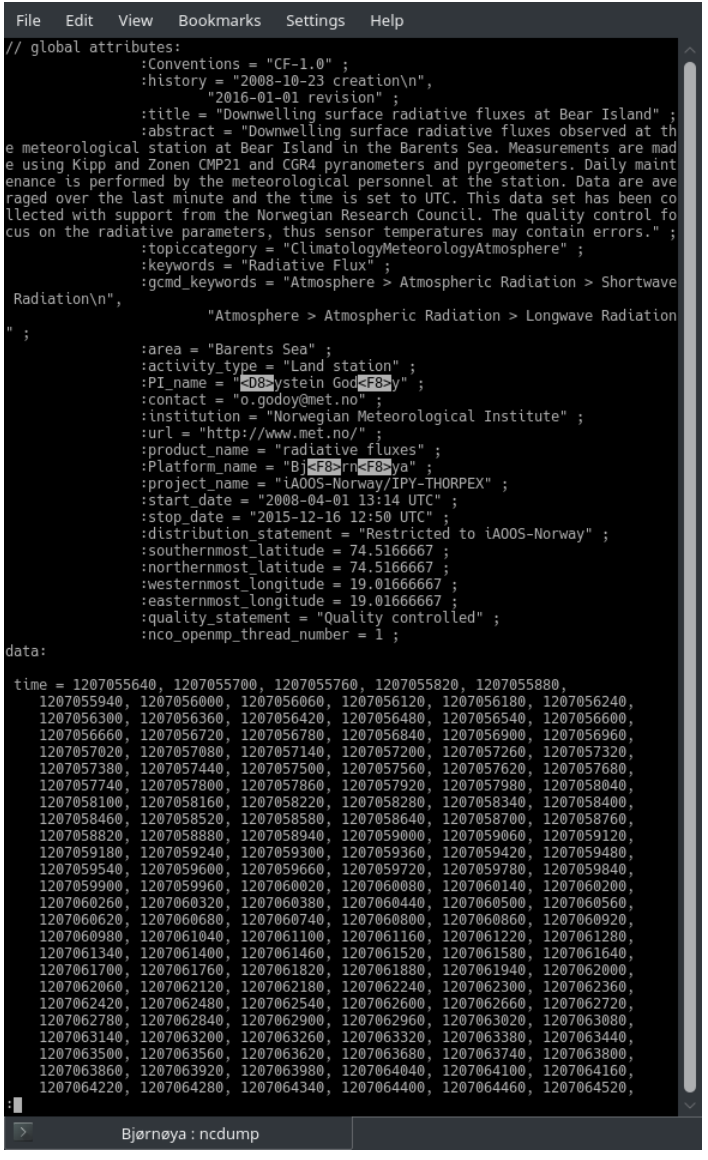


NorDataNet
Norwegian Scientific Data Network

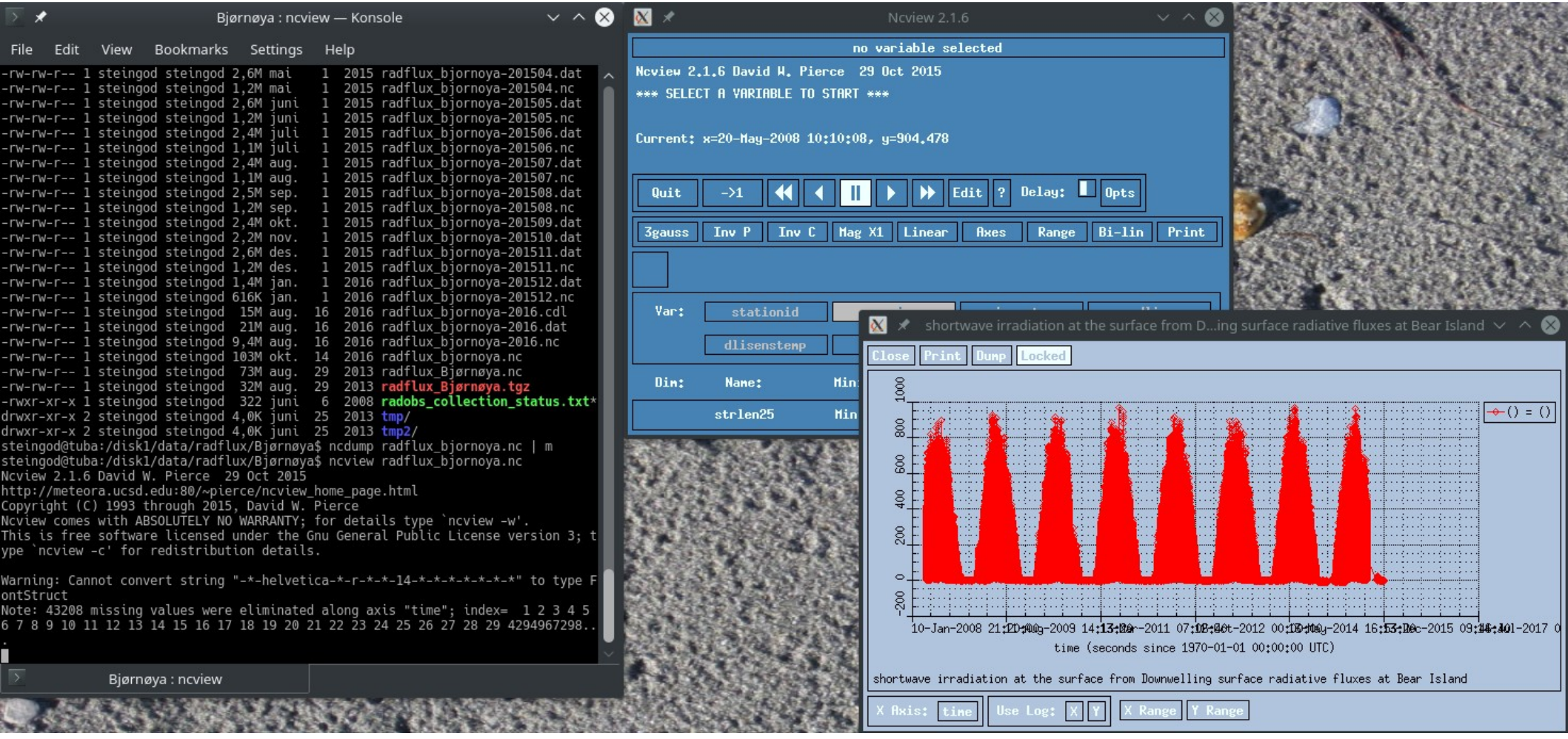


NorDataNet
Norwegian Scientific Data Network

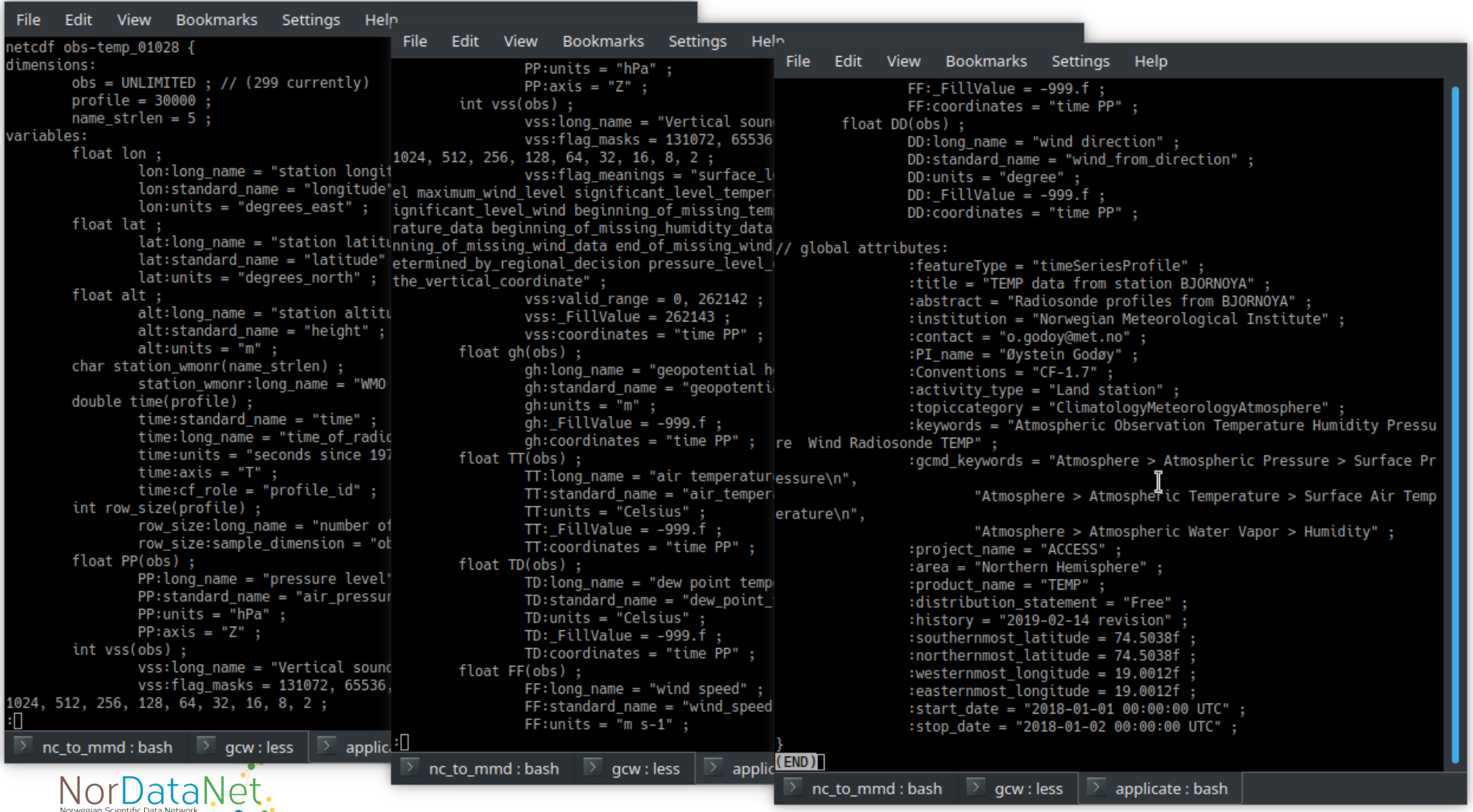




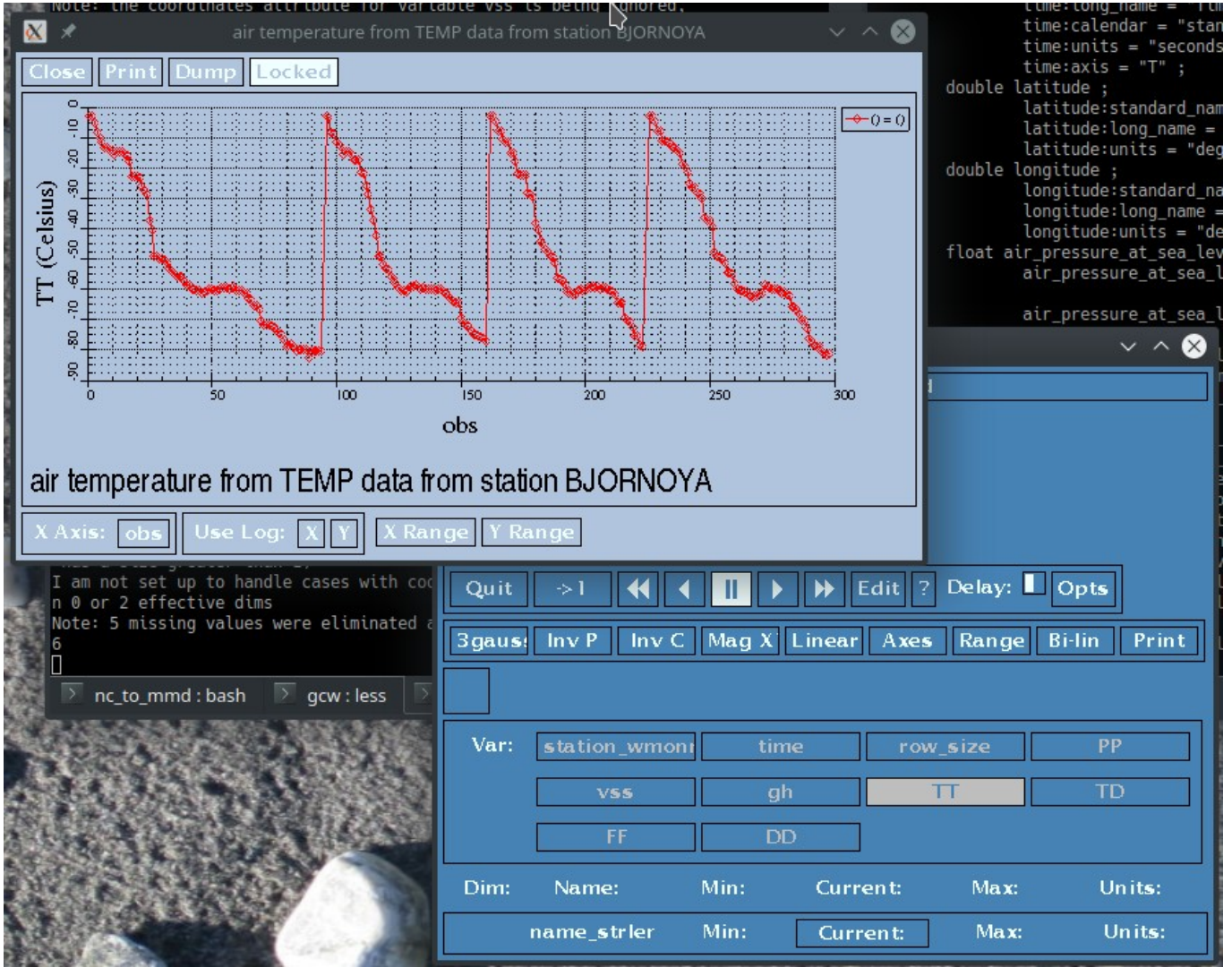
NorDataNet
Norwegian Scientific Data Network



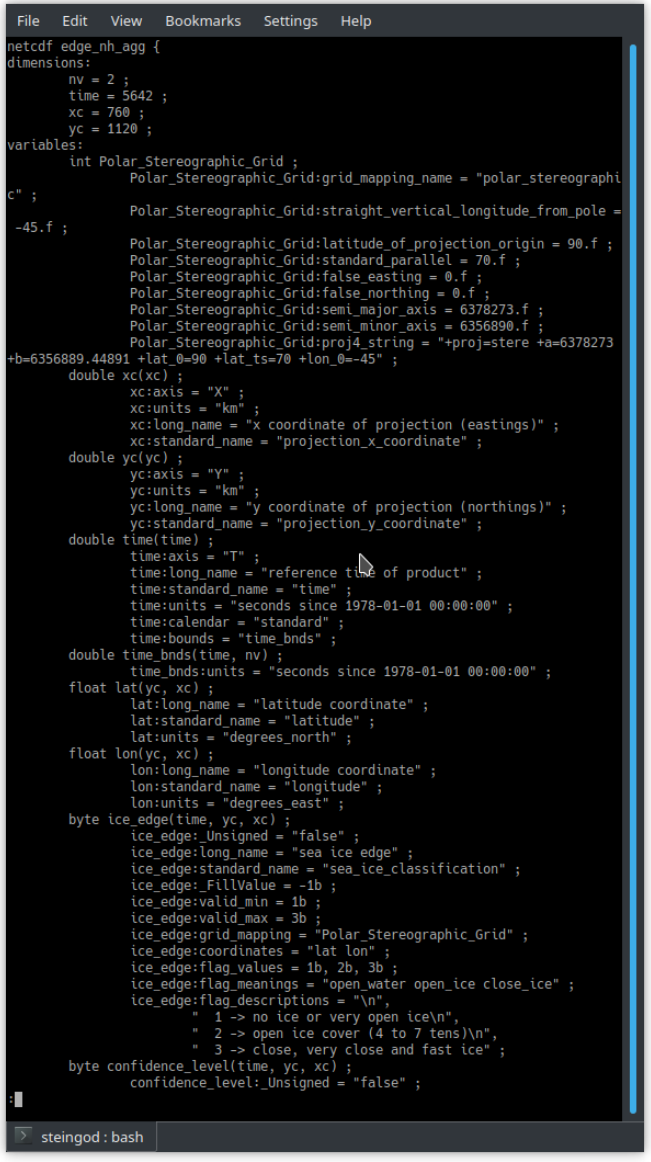
NorDataNet
Norwegian Scientific Data Network

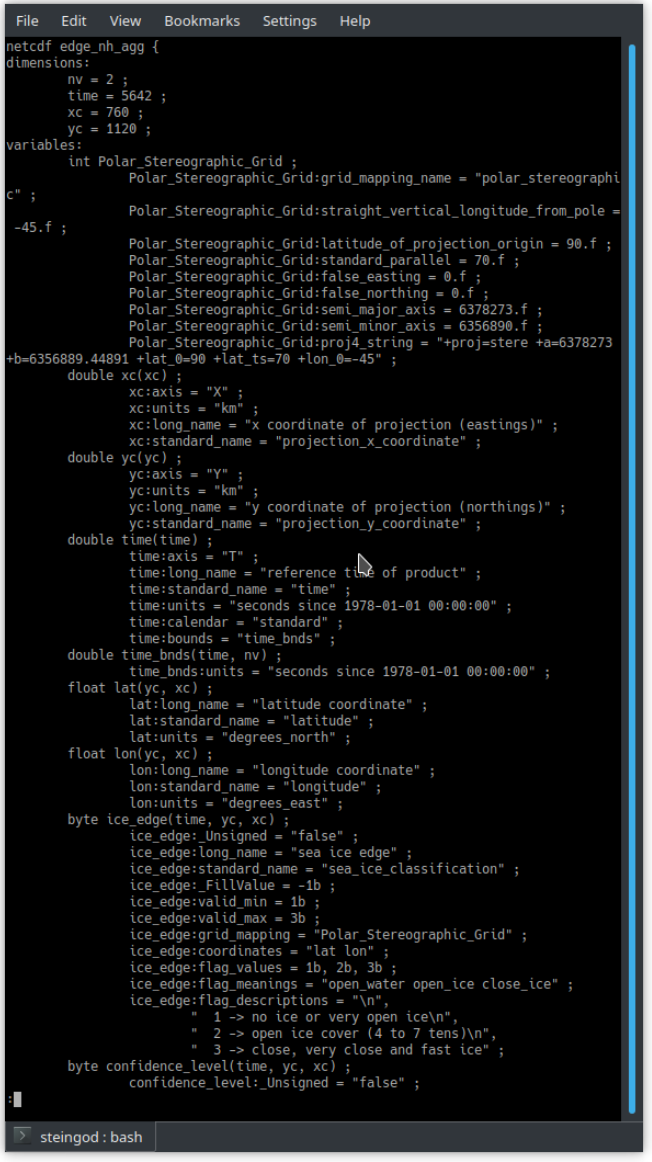


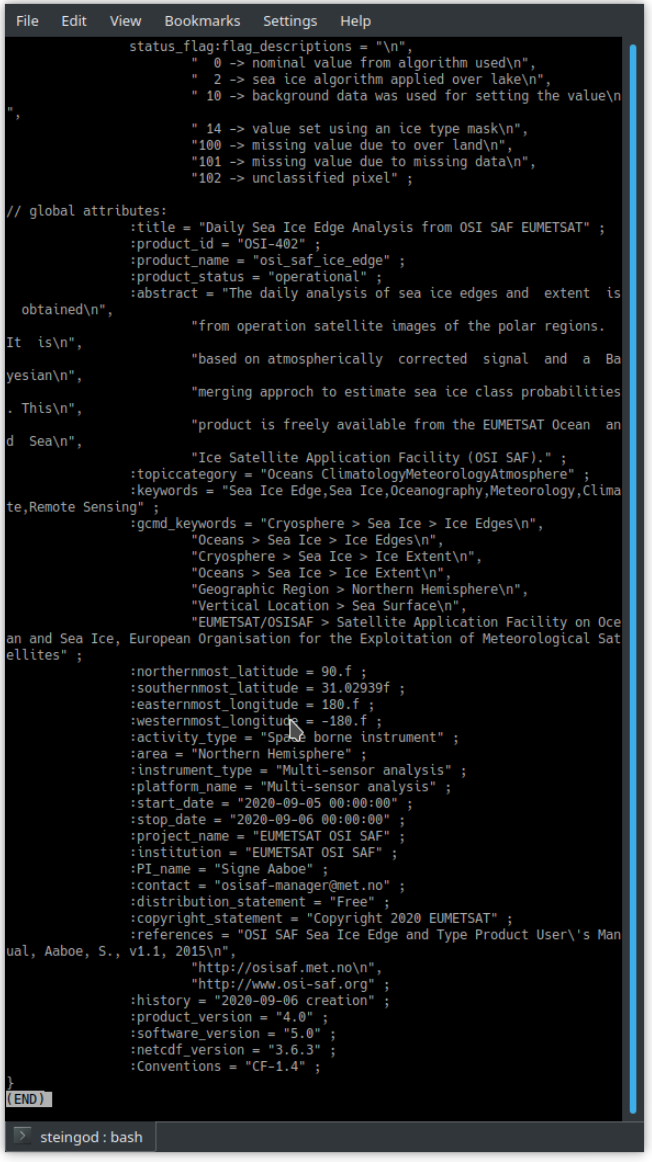
NorDataNet
Norwegian Scientific Data Network

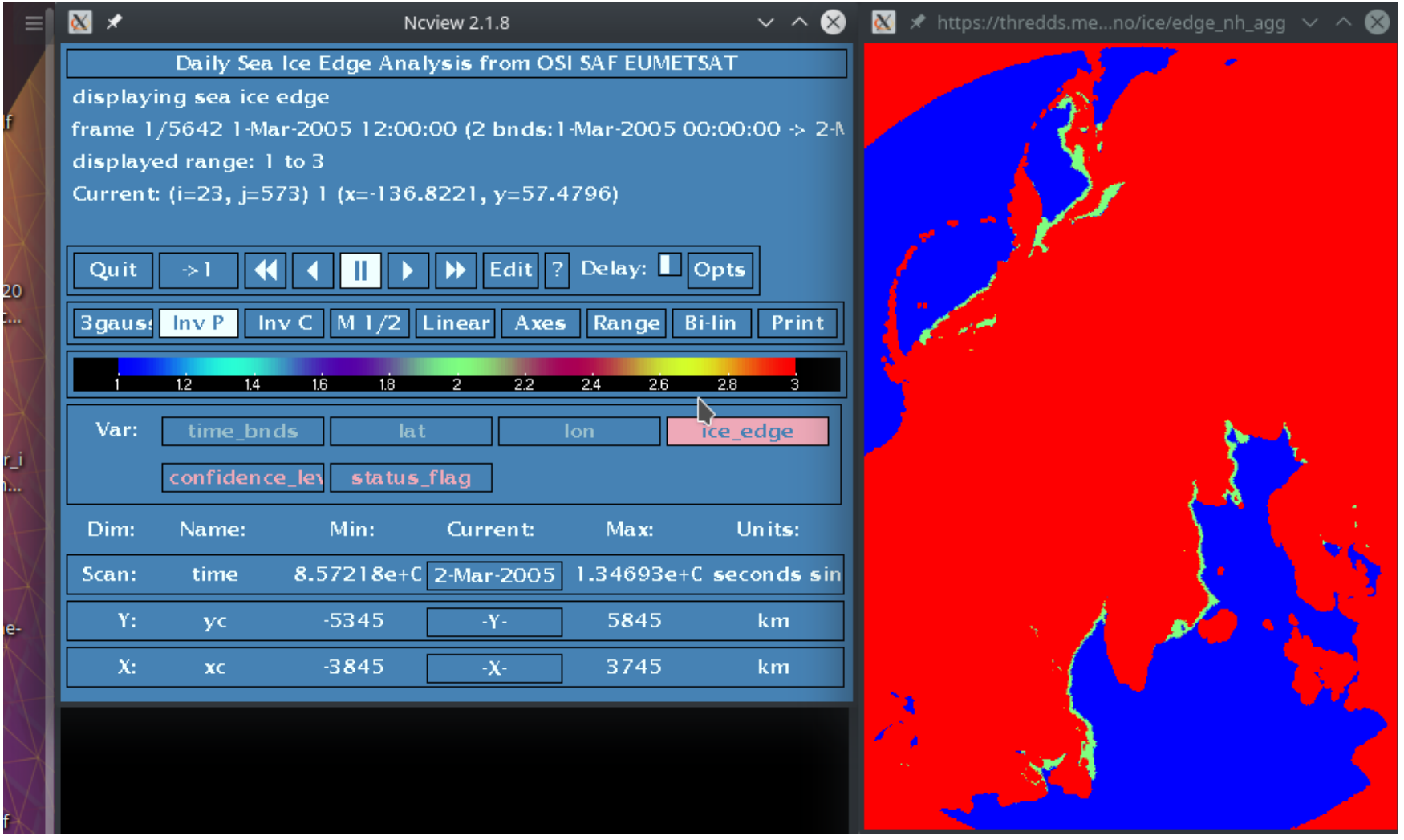


NorDataNet
Norwegian Scientific Data Network


File Edit View Bookmarks Settings Help
netcdf edge_nh_agg {
dimensions:
nv = 2 ;
time = 5642 ;
xc = 760 ;
yc = 1120 ;
variables:
int Polar_Stereographic_Grid ;
Polar_Stereographic_Grid:grid_mapping_name = "polar_stereographi
c" ;
Polar_Stereographic_Grid:straight_vertical_longitude_from_pole =
-45.f ;
Polar_Stereographic_Grid:latitude_of_projection_origin = 90.f ;
Polar_Stereographic_Grid:standard_parallel = 70.f ;
Polar_Stereographic_Grid:false_easting = 0.f ;
Polar_Stereographic_Grid:false_northing = 0.f ;
Polar_Stereographic_Grid:semi_major_axis = 6378273.f ;
Polar_Stereographic_Grid:semi_minor_axis = 6356890.f ;
Polar_Stereographic_Grid:proj4_string = "+proj=stere +a=6378273
+b=6356889.44891 +lat_0=90 +lat_ts=70 +lon_0=-45" ;
double xc(xc) ;
xc:axis = "X" ;
xc:units = "km" ;
xc:long_name = "x coordinate of projection (eastings)" ;
xc:standard_name = "projection_x_coordinate" ;
double yc(yc) ;
yc:axis = "Y" ;
yc:units = "km" ;
yc:long_name = "y coordinate of projection (northings)" ;
yc:standard_name = "projection_y_coordinate" ;
double time(time) ;
time:axis = "T" ;
time:long_name = "reference time of product" ;
time:standard_name = "time" ;
time:units = "seconds since 1978-01-01 00:00:00" ;
time:calendar = "standard" ;
time:bounds = "time_bnds" ;
double time_bnds(time, nv) ;
time_bnds:units = "seconds since 1978-01-01 00:00:00" ;
float lat(yc, xc) ;
lat:long_name = "latitude coordinate" ;
lat:standard_name = "latitude" ;
lat:units = "degrees_north" ;
float lon(yc, xc) ;
lon:long_name = "longitude coordinate" ;
lon:standard_name = "longitude" ;
lon:units = "degrees_east" ;
byte ice_edge(time, yc, xc) ;
ice_edge:_Unsigned = "false" ;
ice_edge:long_name = "sea ice edge" ;
ice_edge:standard_name = "sea_ice_classification" ;
ice_edge:_FillValue = -1b ;
ice_edge:valid_min = 1b ;
ice_edge:valid_max = 3b ;
ice_edge:grid_mapping = "Polar_Stereographic_Grid" ;
ice_edge:coordinates = "lat lon" ;
ice_edge:flag_values = 1b, 2b, 3b ;
ice_edge:flag_meanings = "open_water open_ice close_ice" ;
ice_edge:flag_descriptions = "\n",
" 1 -> no ice or very open ice\n",
" 2 -> open ice cover (4 to 7 tens)\n",
" 3 -> close, very close and fast ice" ;
byte confidence_level(time, yc, xc) ;
confidence_level:_Unsigned = "false" ;
:
steingod : bash


File Edit View Bookmarks Settings Help
netcdf edge_nh_agg {
dimensions:
nv = 2 ;
time = 5642 ;
xc = 760 ;
yc = 1120 ;
variables:
int Polar_Stereographic_Grid ;
Polar_Stereographic_Grid:grid_mapping_name = "polar_stereographi
c" ;
Polar_Stereographic_Grid:straight_vertical_longitude_from_pole =
-45.f ;
Polar_Stereographic_Grid:latitude_of_projection_origin = 90.f ;
Polar_Stereographic_Grid:standard_parallel = 70.f ;
Polar_Stereographic_Grid:false_easting = 0.f ;
Polar_Stereographic_Grid:false_northing = 0.f ;
Polar_Stereographic_Grid:semi_major_axis = 6378273.f ;
Polar_Stereographic_Grid:semi_minor_axis = 6356890.f ;
Polar_Stereographic_Grid:proj4_string = "+proj=stere +a=6378273
+b=6356889.44891 +lat_0=90 +lat_ts=70 +lon_0=-45" ;
double xc(xc) ;
xc:axis = "X" ;
xc:units = "km" ;
xc:long_name = "x coordinate of projection (eastings)" ;
xc:standard_name = "projection_x_coordinate" ;
double yc(yc) ;
yc:axis = "Y" ;
yc:units = "km" ;
yc:long_name = "y coordinate of projection (northings)" ;
yc:standard_name = "projection_y_coordinate" ;
double time(time) ;
time:axis = "T" ;
time:long_name = "reference time of product" ;
time:standard_name = "time" ;
time:units = "seconds since 1978-01-01 00:00:00" ;
time:calendar = "standard" ;
time:bounds = "time_bnds" ;
double time_bnds(time, nv) ;
time_bnds:units = "seconds since 1978-01-01 00:00:00" ;
float lat(yc, xc) ;
lat:long_name = "latitude coordinate" ;
lat:standard_name = "latitude" ;
lat:units = "degrees_north" ;
float lon(yc, xc) ;
lon:long_name = "longitude coordinate" ;
lon:standard_name = "longitude" ;
lon:units = "degrees_east" ;
byte ice_edge(time, yc, xc) ;
ice_edge:_Unsigned = "false" ;
ice_edge:long_name = "sea ice edge" ;
ice_edge:standard_name = "sea_ice_classification" ;
ice_edge:_FillValue = -1b ;
ice_edge:valid_min = 1b ;
ice_edge:valid_max = 3b ;
ice_edge:grid_mapping = "Polar_Stereographic_Grid" ;
ice_edge:coordinates = "lat lon" ;
ice_edge:flag_values = 1b, 2b, 3b ;
ice_edge:flag_meanings = "open_water open_ice close_ice" ;
ice_edge:flag_descriptions = "\n",
" 1 -> no ice or very open ice\n",
" 2 -> open ice cover (4 to 7 tens)\n",
" 3 -> close, very close and fast ice" ;
byte confidence_level(time, yc, xc) ;
confidence_level:_Unsigned = "false" ;
:
steingod : bash

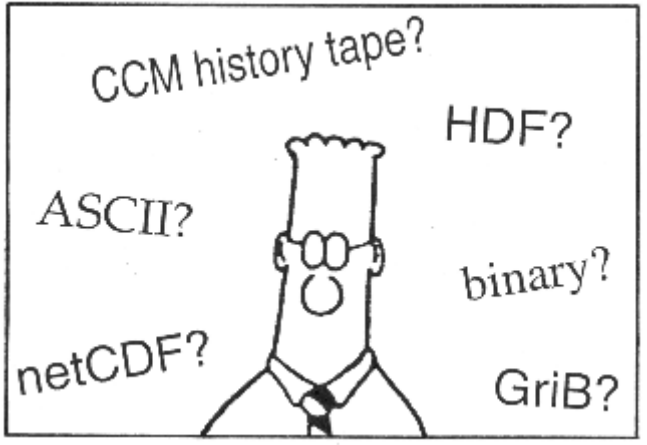

File Edit View Bookmarks Settings Help
status_flag:flag_descriptions = "\n",
" 0 -> nominal value from algorithm used\n",
" 2 -> sea ice algorithm applied over lake\n",
" 10 -> background data was used for setting the value\n
",
" 14 -> value set using an ice type mask\n",
"100 -> missing value due to over land\n",
"101 -> missing value due to missing data\n",
"102 -> unclassified pixel" ;
// global attributes:
:title = "Daily Sea Ice Edge Analysis from OSI SAF EUMETSAT" ;
:product_id = "OSI-402" ;
:product_name = "osi_saf_ice_edge" ;
:product_status = "operational" ;
:abstract = "The daily analysis of sea ice edges and extent is
obtained\n",
"from operation satellite images of the polar regions.
It is\n",
"based on atmospherically corrected signal and a Ba
yesian\n",
"merging approch to estimate sea ice class probabilities
. This\n",
"product is freely available from the EUMETSAT Ocean an
d Sea\n",
"Ice Satellite Application Facility (OSI SAF)." ;
:topiccategory = "Oceans ClimatologyMeteorologyAtmosphere" ;
:keywords = "Sea Ice Edge,Sea Ice,Oceanography,Meteorology,Clima
te,Remote Sensing" ;
:gcmd_keywords = "Cryosphere > Sea Ice > Ice Edges\n",
"Oceans > Sea Ice > Ice Edges\n",
"Cryosphere > Sea Ice > Ice Extent\n",
"Oceans > Sea Ice > Ice Extent\n",
"Geographic Region > Northern Hemisphere\n",
"Vertical Location > Sea Surface\n",
"EUMETSAT/OSISAF > Satellite Application Facility on Oce
an and Sea Ice, European Organisation for the Exploitation of Meteorological Sat
ellites" ;
:northernmost_latitude = 90.f ;
:southernmost_latitude = 31.02939f ;
:easternmost_longitude = 180.f ;
:westernmost_longitude = -180.f ;
:activity_type = "Space borne instrument" ;
:area = "Northern Hemisphere" ;
:instrument_type = "Multi-sensor analysis" ;
:platform_name = "Multi-sensor analysis" ;
:start_date = "2020-09-05 00:00:00" ;
:stop_date = "2020-09-06 00:00:00" ;
:project_name = "EUMETSAT OSI SAF" ;
:institution = "EUMETSAT OSI SAF" ;
:PI_name = "Signe Aaboe" ;
:contact = "osisaf-manager@met.no" ;
:distribution_statement = "Free" ;
:copyright_statement = "Copyright 2020 EUMETSAT" ;
:references = "OSI SAF Sea Ice Edge and Type Product User\'s Man
ual, Aaboe, S., v1.1, 2015\n",
"http://osisaf.met.no\n",
"http://www.osi-saf.org" ;
:history = "2020-09-06 creation" ;
:product_version = "4.0" ;
:software_version = "5.0" ;
:netcdf_version = "3.6.3" ;
:Conventions = "CF-1.4" ;
}
(END)
steingod : bash



NorDataNet
Norwegian Scientific Data Network

Why NetCDF and OPeNDAP?

- NetCDF-CF provides a self describing data
  - in a form compatible with semantic web approaches
  - allowing interdisciplinary use
  - suitable for many different types of data
  - widely used by science communities
  - those not using are often not using standardised approaches
  - Widely supported by analysis tools
  - http://www.cgd.ucar.edu/ccr/bettge/CSM-netCDF/csm_why_netcdf.html
- OPeNDAP provides data to be access over the internet as data streams
  - from programs that weren't originally designed for that purpose,
  - as well as some that were.
  - delivers data, not files
  - Segmenting data in time and space(s)
  - Bridges the message approach of operational data with the long time series approach of climate analysis
  - An OPeNDAP URL might point to an archive containing large volumes of data
  - OPeNDAP provides sophisticated server side sub-sampling capabilities
  - To continue http://docs.opendap.org/index.php/QuickStart



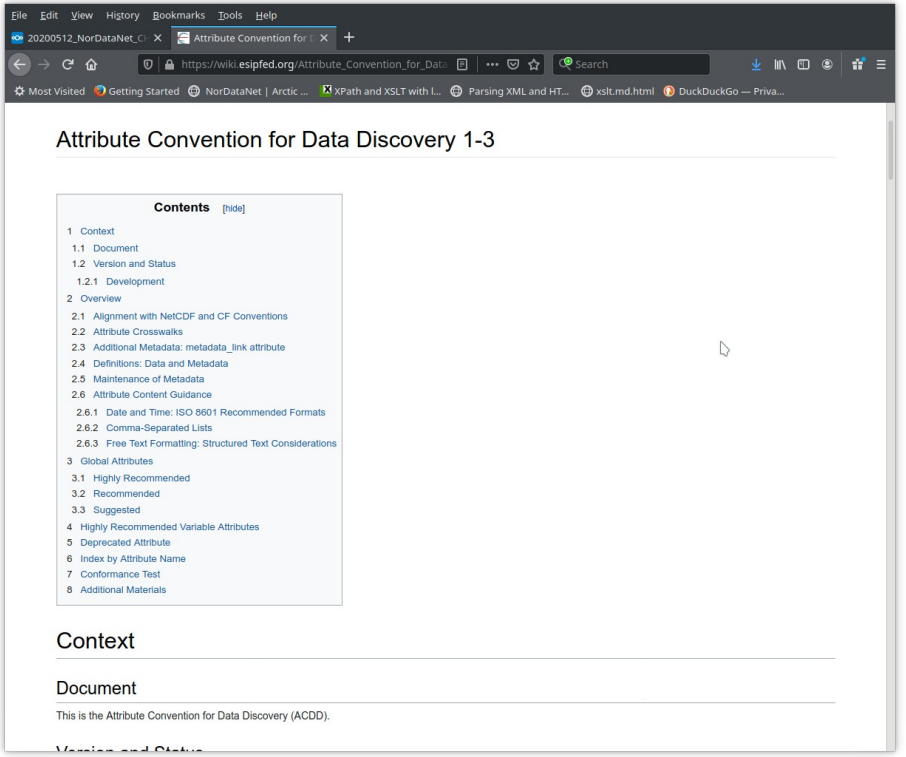
NorDataNet
Norwegian Scientific Data Network

Climate and Forecast Conventions Governance

NorDataNet
Norwegian Scientific Data Network

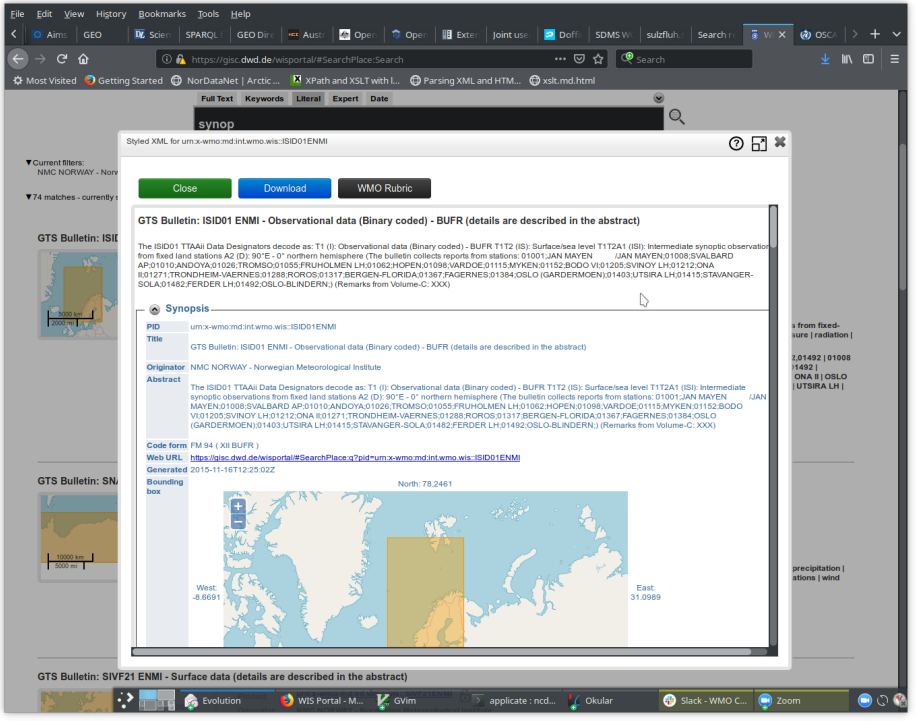
Discovery metadata embedded with data

- Attribute Convention for Data Discovery
- Current in version 1.3
- Allows discovery metadata to be generated automatically
- https://wiki.esipfed.org/Attribute_Convention_for_Data_Discovery_1-3
- https://www.nordatanet.no/en/node/172
- https://adc.met.no/node/4



Granularity

- The level of detail considered in a model or decision making process.
  - The greater the granularity, the more detailed information.
- From granule
  - A small compact particle of substance
- Granular data are detailed data
  - The bits and pieces data are divided into
- To decide on granularity when publishing, think on user perspectives
  - Aggregation can be done automatically



NorDataNet
Norwegian Scientific Data Network