Access NPP files via NcML
(Draft)
October 14, 2010

I. Background

We want to investigate if there is a better way to make NPP files accessible by netCDF-4 classic APIs. Using NcML is one option.

II. What is inside this note

This draft note consists of the following sections:

1) Basic understanding of NcML

2) Is NcML Sufficient?

3) Pros and cons of using NcML

4) Priorities of the NcML features supported by netCDF-4 C library

5) Some NcML examples

III. Basic understanding of NcML

Most contents in this section is from Unidata's NcML tutorial (http://www.unidata.ucar.edu/software/netcdf/ncml/v2.2/Tutorial.html) and Cookbook Examples(http://www.unidata.ucar.edu/software/netcdf/ncml/v2.2/Cookbook.html).

1. Introduction

The NetCDF-Java library can read many different binary file formats, such as netCDF, HDF, GRIB, NEXRAD and read remote datasets through OPeNDAP and other remote access protocols. Any collection of data that can be accessed through the NetCDF-Java library is called a CDM dataset. CDM refers to Common Data Model. A CDM dataset is a file.

The NetCDF Markup Language(NcML) is an XML dialect that allows you to create CDM datasets. An ***NcML document*** is an XML document that uses NcML, and defines a CDM dataset. Commonly, the NcML document refers to another dataset called the ***referenced CDM dataset***.

The purpose of NcML is to allow:

1. Metadata to be added, deleted, and changed.
2. Variables to be renamed, added, deleted and restructured.
3. Data from multiple files (netCDF, HDF etc) to be combined.

The feature of combining multiple files is also called data aggregation. Since data aggregation may not be the focus in this phase, it will not be explained in this document. One can find more information about aggregation at http://www.unidata.ucar.edu/software/netcdf/ncml/v2.2/Aggregation.html

An NcML file should follow the NcML Schema. More information about the schema can be found http://www.unidata.ucar.edu/software/netcdf/ncml/v2.2/AnnotatedSchema4.html.

As we will see in the example, An NcML document can be used to add/delete/change information in the referenced file without changing the file itself.

The following demonstrates how an NcML module works when using an NcML document to add/delete/change information in one referenced file:

It will

1) open the NcML document

2) Find the referenced CDM dataset(file) and open it

3) Retrieve the information about the CDM file structure and corresponding Metadata information

4) Match the add/delete/change information in the NcML document with the retrieved information from the referenced CDM dataset

5) Provide the result requested by applications via netCDF APIs

2. An example to show how to rename/delete/add Metadata

We will illustrate with an example on how one can use NcML to add, rename and delete Metadata.

Given an example netCDF file: Example.nc, the netcdf file contents in CDL format is:

```
netcdf example1 {
dimensions:
        time = UNLIMITED ; // (2 currently)
        lat = 3 ;
        lon = 4 ;
variables:
        int rh(time, lat, lon) ;
                rh:long_name = "relative humidity" ;
                rh:units = "percent" ;
        double T(time, lat, lon) ;
                T:long_name = "surface temperature" ;
                T:units = "degC" ;
        float lat(lat) ;
                lat:units = "degrees_north" ;
        float lon(lon) ;
                lon:units = "degrees_east" ;
        int time(time) ;
                time:units = "hours" ;
// global attributes:
                :title = "Example Data" ;
}
```

We want to use an NcML document to

1) Change the value of a global attribute named “title”

2) Add a global attribute named “Conventions”

3) Rename variable “rh” to “RelativeHumidity”

4) Add a variable(RelativeHumidty) attribute “Standard_name”

5) Add a variable(T) attribute “Standard_name”

6) Change a variable(T) attribute “units” value

7) A new variable is “deltaLat” added. Since it doesn’t exist in the referenced file, the value must be defined.

The NcML document is as follows with the step index marked.

```
<?xml version="1.0" encoding="UTF-8"?>
<netcdf xmlns="http://www.unidata.ucar.edu/namespaces/netcdf/ncml-2.2"
location="example1.nc">

(1)<attribute name="title" type="String" value="Example Data using CF" />
(2)<attribute name="Conventions" value="CF-1.0" />

(3)<variable name="RelativeHumidity" orgName="rh">
(4)  <attribute name="standard_name" type="String" value="relative humidity" />
   </variable>

   <variable name="T">
(5)  <attribute name="standard_name" type="String" value="temperature" />
(6)  <attribute name="units" type="String" value="degreesC" />
   </variable>


(7)<variable name="deltaLat" type="double" shape="lat">
     <values>.1 .1 .01</values>
   </variable>

</netcdf>
```

The ncdump (implemented by netCDF Java) can then dump the netCDF dataset with added/changed information as follows.

```
netcdf file:C:/temp/exercise3.ncml {
  dimensions:
   time = UNLIMITED;   // (2 currently)
   lat = 3;
   lon = 4;

  variables:
   int RelativeHumidity(time=2, lat=3, lon=4);
     :long_name = "relative humidity";
     :units = "percent";
     :standard_name = "relative humidity";
   double T(time=2, lat=3, lon=4);
     :long_name = "surface temperature";
     :units = "degreesC";
     :standard_name = "temperature";
   float lat(lat=3);
     :units = "degrees_north";
   float lon(lon=4);
     :units = "degrees_east";
   int time(time=2);
     :units = "hours";
   double deltaLat(lat=3);


 :title = "Example Data using CF";
 :Conventions = "CF-1.0";
}
```

Note one just needs to provide the NcML document name to obtain the above results.

IV. Is NcML sufficient

Based on the current NPP XML file(check the NPP XML note), it seems that at least dimension name,length pair for each HDF5 dataset, dataset attributes “Units”, “fill value”(several of them), “scale”, “offset” need to pass to netCDF-4 APIs so that end users can easily understand the meaning of the data. Key CF attributes “coordinates” for normal variables and “units” for coordinate variables should be added if these information are available from either the NPP HDF5 file or from other NPP documents.

According to NcML schema (http://www.unidata.ucar.edu/software/netcdf/ncml/v2.2/AnnotatedSchema4.html), all the above features are supported.

Moreover, one can use NcML to rename variables and attributes. This feature makes it easy to rename attributes containing special characters.

NcML also supports the “group” concept. The current NetCDF Java package can successfully generate An NcML file from an HDF-EOS5 file that have multiple HDF5 groups. See the Appendix 2 for details. There should be no problems if we need to add/delete/change some attributes under an HDF5 group.

Based on the current understanding, NcML is sufficient for our purpose.

V. Pros and cons of using NcML

Pros:

1. NcML provides a way to modify the NPP metadata or even data by not touching NPP HDF5 files.

2. NcML schema is well established and starts to be used operationally. Unidata's THREEDS catalog server and OPeNDAP already include NcML modules. A google search identifies a NOAA data center also uses NcML. It is not clear if the NOAA data center uses NcML operationally or for research only, though. It seems to be very possible that there will be more NcML users by next November.

3. NetCDF Java library already implemented the NcML module. Since NcML Java module starts to be widely accepted by the user communities and it may be easier to implement the NcML module in netCDF C APIs based on Java NcML module, the priority for Unidata to implement the NcML module in netCDF C library is much higher and in consequence, the chance for the feature released in first-half of the next year may be bigger.

Cons:

It is uncertain when the NcML module in netCDF-C library will be available. One big factor of the success of this project depends on the availability and the robustness of the NcML module as well as other features in the next netCDF-4 release or the release before the first half of year 2011.

Currently netCDF Java fails to generate the NcML file of the NPP sample file. This may or may not be a big problem though.

VI. Priorities of the NcML features supported by netCDF-4 C library

1. Add dimension name and length pair

2. Add/rename attributes

3. Add variables

4. Rename variable names

5. Add/delete/change groups - this may depend on the new features of netCDF-4 classic APIs

6. Various aggregation features – may not be necessary in this phase

Appendix

1. An NcML document that is used to add metadata to a netCDF file

This NcML file is adapted from a NOAA’s NcML application. Note that _FillValue and coordinates attributes may also be added to NPP files.

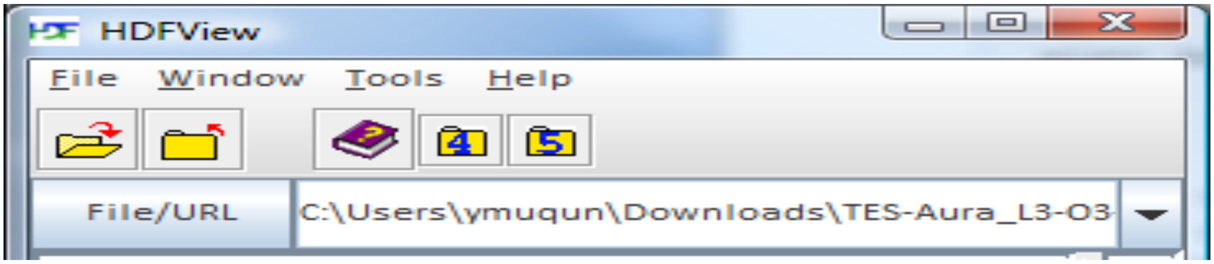
```
<?xml version="1.0" encoding="UTF-8"?>
<netcdf xmlns=http://www.unidata.ucar.edu/namespaces/netcdf/ncml-2.2
location=”ecom_lonlat.nc”>

      <variable name="elev">
        <attribute name="standard_name" value="sea_surface_height"/>
        <!-- for each variable, need to add the "coordinates" attribute,
pointing to coordinate variables "lon lat"-->
        <attribute name="coordinates" type="String" value="lon lat"/>
        <attribute name="_FillValue" type="short" value="0"/>
      </variable>
      <variable name="heat_flux">
        <attribute name="coordinates" type="String" value="lon lat"/>
      </variable>
      <variable name="cd">
        <attribute name="coordinates" type="String" value="lon lat"/>
      </variable>
      <variable name="depth">
        <attribute name="coordinates" type="String" value="lon lat"/>
        <attribute name="_FillValue" type="short" value="-99999"/>
      </variable>
      <variable name="temp">
        <attribute name="coordinates" type="String" value="lon lat zpos"/>
        <attribute name="_FillValue" type="short" value="-23405"/>
      </variable>
      <variable name="salt">
        <attribute name="coordinates" type="String" value="lon lat zpos"/>
        <attribute name="_FillValue" type="short" value="-32767"/>
      </variable>
      <variable name="conc">
        <attribute name="coordinates" type="String" value="lon lat zpos"/>
      </variable>
      <variable name="u">
        <attribute name="coordinates" type="String" value="lon lat zpos"/>
        <remove type="attribute" name="long_name"/>
        <attribute name="standard_name"
value="x_grid_directed_sea_water_velocity"/>
        <attribute name="_FillValue" type="short" value="0"/>
      </variable>
      <variable name="v">
        <attribute name="coordinates" type="String" value="lon lat zpos"/>
        <remove type="attribute" name="long_name"/>
        <attribute name="standard_name"
value="y_grid_directed_sea_water_velocity"/>
        <attribute name="_FillValue" type="short" value="0"/>
      </variable>
      <variable name="zpos" orgName="sigma">
        <attribute name="standard_name" type="String"
value="ocean_sigma_coordinate"/>
        <attribute name="formula_terms" type="String" value="sigma: zpos eta:
elev depth: depth"/>
        <attribute name="positive" type="String" value="up"/>
        <attribute name="axis" type="String" value="Z"/>
        <attribute name="units" type="String" value="1"/>
      </variable>
      <attribute name="Conventions" type="String" value="CF-1.0"/>
</netcdf>
```

2. An NcML representation of an HDF-EOS5 file

This file is an HDF-EOS5 grid file. It has several nested groups, dozens of group attributes and dozens of HDF5 datasets.

1). Where to find the file: ftp://ftp.hdfgroup.uiuc.edu/pub/outgoing/eseo2/aura_samples/grid/TES-Aura_L3-O3-M2008m03_F01_04.he5

2). Screenshot of the HDF-EOS5 file



C:\Users\ymuqun\Downloads\TES-Aura_L3-O3

3). NcML representation of the HDF-EOS5 file

```
<?xml version="1.0" encoding="UTF-8"?>
```

```
<netcdf xmlns="http://www.unidata.ucar.edu/namespaces/netcdf/ncml-2.2"
location="C:/Users/kent/Downloads/TES-Aura_L3-O3-M2008m03_F01_04.he5">
  <group name="HDFEOS">
    <group name="ADDITIONAL">
      <group name="FILE_ATTRIBUTES">
        <attribute name="InstrumentName" value="TES" />
        <attribute name="ProcessLevel" value="L3" />
        <attribute name="OrbitNumber" type="int" value="19310 19734" />
        <attribute name="OrbitPeriod" type="double" value="-999.0 -999.0" />
        <attribute name="TAI93At0zOfGranule" type="double" value="4.78600823222808E8" />
        <attribute name="GlobalSurveyNumber"
value="6628,6633,6641,6646,6651,6652,6654,6656,6658,6663,6668,6679,6708,6734,6757" />
        <attribute name="GranuleMonth" type="byte" value="3" />
        <attribute name="GranuleDay" type="byte" value="2" />
        <attribute name="GranuleYear" type="int" value="2008" />
        <attribute name="GranuleDayOfYear" type="short" value="62" />
        <attribute name="SurveyMode" value="Global" />
        <attribute name="PGEVersion" value="R10.02.00" />
        <attribute name="StartUTC" value="2008-03-02T08:40:17.222808Z" />
        <attribute name="EndUTC" value="2008-03-31T11:23:38.980407Z" />
        <attribute name="Period" value="Monthly" />
        <attribute name="Command_Seq_ID" type="int" value="41" />
      </group>
    </group>
    <group name="GRIDS">
      <group name="NadirGrid">
        <attribute name="Projection" value="Geographic" />
        <attribute name="GridOrigin" value="Center" />
        <attribute name="GridSpacing" value="(4,2)" />
        <attribute name="GridSpacingUnit" value="deg" />
        <attribute name="GridSpan" value="(0,360,-82,+82)" />
        <attribute name="GridSpanUnit" value="deg" />
        <attribute name="DeltaFullLatitude" type="float" value="4.0" />
        <attribute name="DeltaFullLongitude" type="float" value="8.0" />
        <attribute name="MonthlyL3Algorithm" value="Binning average, L2 data weighted inversely by
error and distance to L3 geolocation point" />
        <group name="Data_Fields">
          <variable name="Latitude" shape="83" type="float">
            <attribute name="_FillValue" type="float" value="-999.0" />
            <attribute name="MissingValue" type="float" value="-999.0" />
            <attribute name="Title" value="Latitude" />
            <attribute name="Units" value="deg" />
            <attribute name="UniqueFieldDefinition" value="HIRDLS-MLS-TES-SHARED" />
            <attribute name="_LastModified" value="2008-05-29T01:25:41Z" />
          </variable>
          <variable name="Longitude" shape="90" type="float">
            <attribute name="_FillValue" type="float" value="-999.0" />
            <attribute name="MissingValue" type="float" value="-999.0" />
      <attribute name="Title" value="Longitude" />
      <attribute name="Units" value="deg" />
      <attribute name="UniqueFieldDefinition" value="HIRDLS-MLS-TES-SHARED" />
      <attribute name="_LastModified" value="2008-05-29T01:25:41Z" />
    </variable>
    <variable name="O3" shape="90 83 15" type="float">
      <attribute name="MissingValue" type="float" value="-999.0" />
      <attribute name="Title" value="O3" />
      <attribute name="Units" value="VMR" />
      <attribute name="UniqueFieldDefinition" value="HIRDLS-MLS-TES-SHARED" />
      <attribute name="_FillValue" type="float" value="-999.0" />
      <attribute name="_LastModified" value="2008-05-29T01:25:41Z" />
    </variable>
    <variable name="O3DataCount" shape="90 83 15" type="short">
      <attribute name="_FillValue" type="short" value="-999" />
      <attribute name="MissingValue" type="short" value="-999" />
      <attribute name="Title" value="O3DataCount" />
      <attribute name="Units" value="N/A" />
      <attribute name="UniqueFieldDefinition" value="TES-SPECIFIC" />
      <attribute name="_LastModified" value="2008-05-29T01:25:41Z" />
    </variable>
    <variable name="O3Maximum" shape="90 83 15" type="float">
      <attribute name="MissingValue" type="float" value="-999.0" />
      <attribute name="Title" value="O3Maximum" />
      <attribute name="Units" value="VMR" />
      <attribute name="UniqueFieldDefinition" value="TES-SPECIFIC" />
      <attribute name="_FillValue" type="float" value="-999.0" />
      <attribute name="_LastModified" value="2008-05-29T01:25:41Z" />
    </variable>
    <variable name="O3Minimum" shape="90 83 15" type="float">
      <attribute name="MissingValue" type="float" value="-999.0" />
      <attribute name="Title" value="O3Minimum" />
      <attribute name="Units" value="VMR" />
      <attribute name="UniqueFieldDefinition" value="TES-SPECIFIC" />
      <attribute name="_FillValue" type="float" value="-999.0" />
      <attribute name="_LastModified" value="2008-05-29T01:25:41Z" />
    </variable>
    <variable name="O3StdDeviation" shape="90 83 15" type="float">
      <attribute name="MissingValue" type="float" value="-999.0" />
      <attribute name="Title" value="O3StdDeviation" />
      <attribute name="Units" value="N/A" />
      <attribute name="UniqueFieldDefinition" value="TES-SPECIFIC" />
      <attribute name="_FillValue" type="float" value="-999.0" />
      <attribute name="_LastModified" value="2008-05-29T01:25:41Z" />
    </variable>
    <variable name="OzoneTropColumn" shape="90 83" type="float">
      <attribute name="MissingValue" type="float" value="-999.0" />
      <attribute name="Title" value="OzoneTropColumn" />
      <attribute name="Units" value="Molecules/cm^2" />
      <attribute name="UniqueFieldDefinition" value="TES-SPECIFIC" />
      <attribute name="_FillValue" type="float" value="-999.0" />
      <attribute name="_LastModified" value="2008-05-29T01:25:41Z" />
    </variable>
    <variable name="Pressure" shape="15" type="float">
      <attribute name="_FillValue" type="float" value="-999.0" />
      <attribute name="MissingValue" type="float" value="-999.0" />
      <attribute name="Title" value="Pressure" />
      <attribute name="Units" value="hPa" />
      <attribute name="UniqueFieldDefinition" value="HIRDLS-MLS-TES-SHARED" />
      <attribute name="_LastModified" value="2008-05-29T01:25:41Z" />
    </variable>
    <variable name="TotColDensDataCount" shape="90 83" type="short">
      <attribute name="_FillValue" type="short" value="-999" />
      <attribute name="MissingValue" type="short" value="-999" />
      <attribute name="Title" value="TotColDensDataCount" />
      <attribute name="Units" value="N/A" />
      <attribute name="UniqueFieldDefinition" value="TES-SPECIFIC" />
      <attribute name="_LastModified" value="2008-05-29T01:25:41Z" />
    </variable>
    <variable name="TotColDensMaximum" shape="90 83" type="float">
      <attribute name="MissingValue" type="float" value="-999.0" />
      <attribute name="Title" value="TotColDensMaximum" />
      <attribute name="Units" value="Molecules/cm^2" />
      <attribute name="UniqueFieldDefinition" value="TES-SPECIFIC" />
      <attribute name="_FillValue" type="float" value="-999.0" />
      <attribute name="_LastModified" value="2008-05-29T01:25:41Z" />
    </variable>
    <variable name="TotColDensMinimum" shape="90 83" type="float">
      <attribute name="MissingValue" type="float" value="-999.0" />
      <attribute name="Title" value="TotColDensMinimum" />
      <attribute name="Units" value="Molecules/cm^2" />
      <attribute name="UniqueFieldDefinition" value="TES-SPECIFIC" />
      <attribute name="_FillValue" type="float" value="-999.0" />
      <attribute name="_LastModified" value="2008-05-29T01:25:41Z" />
    </variable>
    <variable name="TotColDensStdDeviation" shape="90 83" type="float">
      <attribute name="MissingValue" type="float" value="-999.0" />
      <attribute name="Title" value="TotColDensStdDeviation" />
      <attribute name="Units" value="N/A" />
      <attribute name="UniqueFieldDefinition" value="TES-SPECIFIC" />
      <attribute name="_FillValue" type="float" value="-999.0" />
      <attribute name="_LastModified" value="2008-05-29T01:25:41Z" />
    </variable>
    <variable name="TotalColumnDensity" shape="90 83" type="float">
      <attribute name="MissingValue" type="float" value="-999.0" />
      <attribute name="Title" value="TotalColumnDensity" />
      <attribute name="Units" value="Molecules/cm^2" />
      <attribute name="UniqueFieldDefinition" value="TES-SPECIFIC" />
      <attribute name="_FillValue" type="float" value="-999.0" />
      <attribute name="_LastModified" value="2008-05-29T01:25:41Z" />
     </variable>
    </group>
   </group>
  </group>
 </group>
 <group name="HDFEOS_INFORMATION">
  <attribute name="HDFEOSVersion" value="HDFEOS_5.1.9" />
  <variable name="StructMetadata.0" shape="32000" type="char">
   <attribute name="_LastModified" value="2008-05-29T01:25:41Z" />
  </variable>
  <variable name="coremetadata" shape="65535" type="char">
   <attribute name="_LastModified" value="2008-05-29T01:25:41Z" />
  </variable>
 </group>
</netcdf>
```