iPres 02010

JHOVE2

JaNeME

J-NetCDF Metadata Extractor

José A. Mejía V., M.SC.

14.09.2010

WissGrid

HELMHOLTZ GEMEINSCHAFT

AWI

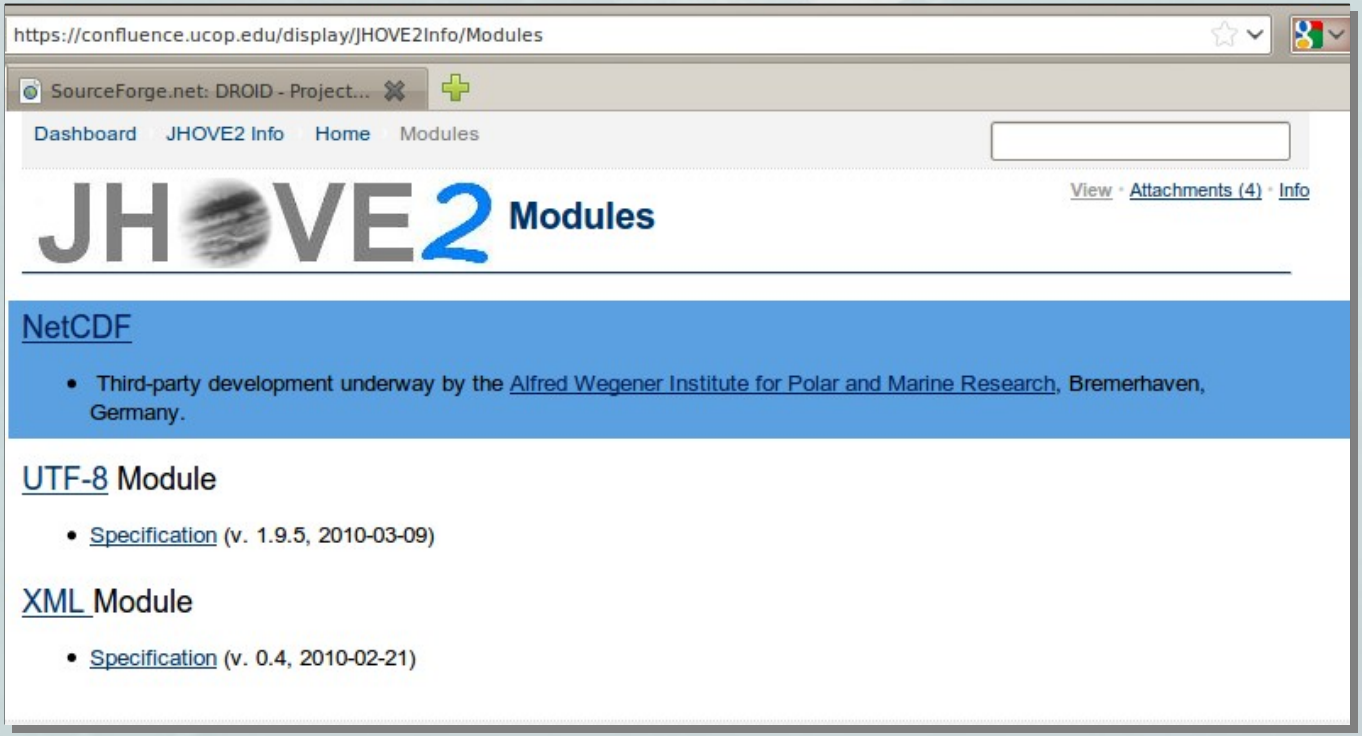
Introduction to JaNeME

The beginning of knowledge is the discovery of something we do not understand.

Frank Herbert (1920 - 1986)

Introduction to JaNeME

- The NetCDF and GRIB Formatmodules have been developed at the Alfred Wegener Institute for Polar and Marine Research [awi] in scope of the work package # 3 – Long Term Preservation of Research Archives – of Wissgrid [lza]



Introduction to JaNeME

- BSD License
- Based on Jhove2 0.6.0
- Parsing und Characterization of files in NetCDF 3.0/4.0 and Grib 1.0/2.0 formats via the Java-netcdf Library 4.1 [jnc]
- Available for free download at http://aforge.awi.de/gf/project/jhove2/frs

Introduction to JaNeME

•The native XML, JSON, and Text outputs are also supported!

NetCDF & GRIB File Formats

As a general rule the most successful man in life is the man who has the best information.

Benjamin Disraeli (1804 - 1881)

NetCDF Format

Network Common Data Form [nc]

File Extension
.nc .cdf

Mediatype
application/netcdf
application/x-netcdf

Developed by
UCAR (University Corporation for Atmospheric Research)

Formattype
Scientific Binary File

- Open Standard frequently used in climate research, oceanography, meteorology and GIS (geographic information systems).
- Self describing binary format, whose header provides with metadata and information about the structure of the data section.
- Its metadata consists of a map with pairs *"Key & Value"*
- The data content is stored as *n*-dimensional arrays.

NetCDF Format

NetCDF Versions Supported by JaNeME

JaNEME's Formatprofile	Format	Signature (ASCII C Notation)
NetCDF3Profile	Classic format	CDF \001
	64-bit offset format	CDF \002
NetCDF4Profile	NetCDF-4 based on HDF5	\211 HDF \r \n \032 \n

GRIB Format

GRIdded Binary [gb]

File Extension	.grb .wmo
Specifications	[grib-FM92-IX] [grib-FM92-XII]
Developed by	World Meteorologial Organization WMO CBS
Formattype	Scientific Binary File

- A mathematically concise data format commonly used in meteorology to store historical and forecast weather data.
- Standardized by the World Meteorological Organization's Commission for Basic Systems, known under number GRIB FM 92-IX.
- Most GRIB files are actually a collection of individual self-containing records, and the individual records itself can stand alone as meaningful data.
- Each individual GRIB record has two components - the part that describes the record (the header), and the actual binary data itself.

GRIB Format

GRIdded Binary [gb]

File Extension	.grb .wmo
Specifications	[grib-FM92-IX] [grib-FM92-XII]
Developed by	World Meteorologial Organization WMO CBS
Formattype	Scientific Binary File

•There are 2 parts of the GRIB 1 header - one mandatory (Product Definition Section - PDS) and one optional (Grid Description Section – GDS).

•PDS describe who created the data (the research/operation center), the involved numerical model/process, the data that is actually stored (such as wind, temperature, ozone concentration etc.), units of the data (meters, pressure etc.), vertical system of the data (constant height, constant pressure, constant potential temperature), and the time stamp.

GRIB Format

GRIB Versions Supported by JaNeME

JaNEME's Formatprofile	Format	Signature (ASCII C Notation)
GRIB1Profile	GRIB 1	GRIB 0xhh 0xhh 0xhh 0x01
GRIB2Profile	GRIB 2	GRIB 0xhh 0xhh 0xhh 0x02

DROID Configuration

- NetCDF and GRIB files are not included in the [Pronom] database.

- Generate a new signature file, invoking the following Ed script

(cat script && echo w) | ed - DROID_Signature_V20.xml

```
8439c
        <FileFormat ID="9001" Name="NetCDF-3 classic and large format"
            Version="1.0" PUID="application/netcdf-3">
            <Extension>nc</Extension>
            <Extension>cdf</Extension>
            <InternalSignatureID>9991</InternalSignatureID>
        </FileFormat>
        <FileFormat ID="9002" Name="NetCDF-3 classic and large format 64-bit"
            Version="1.0" PUID="application/netcdf-3">
            <Extension>nc</Extension>
            <Extension>cdf</Extension>
            <InternalSignatureID>9992</InternalSignatureID>
        </FileFormat>
        <FileFormat ID="9003" Name="NetCDF-4 built on HDF-5"
            Version="1.0" PUID="application/netcdf-4">
            <Extension>nc</Extension>
            <Extension>cdf</Extension>
            <InternalSignatureID>9993</InternalSignatureID>
        </FileFormat>
        <FileFormat ID="9004" Name="WMO Grib Edition 1" Version="1.0"
            PUID="grib-FM92-IX">
            <Extension>grb</Extension>
            <Extension>wmo</Extension>
            <InternalSignatureID>9994</InternalSignatureID>
        </FileFormat>
        <FileFormat ID="9005" Name="WMO Grib Edition 2" Version="1.0"
            PUID="grib-FM92-XII">
            <Extension>grb</Extension>
            <Extension>wmo</Extension>
            <InternalSignatureID>9995</InternalSignatureID>
        </FileFormat>
    </FileFormatCollection>
.
```

```
6326a
        <InternalSignature ID="9991" Specificity="Specific">
            <ByteSequence Reference="BOFoffset">
                <SubSequence Position="1" SubSeqMinOffset="0"
                    SubSeqMaxOffset="0" MinFragLength="0">
                    <Sequence>43444601</Sequence>
                    <DefaultShift>5</DefaultShift>
                    <Shift Byte="01">1</Shift>    <Shift Byte="46">2</Shift>
                    <Shift Byte="44">3</Shift>    <Shift Byte="43">4</Shift>
                </SubSequence>
            </ByteSequence>
        </InternalSignature>
        <InternalSignature ID="9992" Specificity="Specific">
            <ByteSequence Reference="BOFoffset">
                <SubSequence Position="1" SubSeqMinOffset="0"
                    SubSeqMaxOffset="0" MinFragLength="0">
                    <Sequence>43444602</Sequence>
                    <DefaultShift>5</DefaultShift>
                    <Shift Byte="02">1</Shift>    <Shift Byte="46">2</Shift>
                    <Shift Byte="44">3</Shift>    <Shift Byte="43">4</Shift>
                </SubSequence>
            </ByteSequence>
        </InternalSignature>
        <InternalSignature ID="9993" Specificity="Specific">
            <ByteSequence Reference="BOFoffset">
                <SubSequence Position="1" SubSeqMinOffset="0"
                    SubSeqMaxOffset="0" MinFragLength="0">
                    <Sequence>894844460d0a1a0a</Sequence>
                    <DefaultShift>9</DefaultShift>
                    <Shift Byte="0a">1</Shift>    <Shift Byte="1a">2</Shift>
                    <Shift Byte="0d">4</Shift>    <Shift Byte="46">5</Shift>
                    <Shift Byte="44">6</Shift>    <Shift Byte="48">7</Shift>
                    <Shift Byte="89">8</Shift>    </SubSequence>
            </ByteSequence>
        </InternalSignature>
```

```
<InternalSignature ID="9994" Specificity="Specific">
    <ByteSequence Reference="BOFoffset">
        <SubSequence Position="1" SubSeqMinOffset="0"
            SubSeqMaxOffset="0" MinFragLength="0">
            <Sequence>47524942</Sequence>
            <DefaultShift>5</DefaultShift>
            <Shift Byte="42">1</Shift>
            <Shift Byte="49">2</Shift>
            <Shift Byte="52">3</Shift>
            <Shift Byte="47">4</Shift>
            <RightFragment Position="1" MinOffset="3" MaxOffset="3">01</RightFragment>
        </SubSequence>
    </ByteSequence>
</InternalSignature>
<InternalSignature ID="9995" Specificity="Specific">
    <ByteSequence Reference="BOFoffset">
        <SubSequence Position="1" SubSeqMinOffset="0"
            SubSeqMaxOffset="0" MinFragLength="0">
            <Sequence>47524942</Sequence>
            <DefaultShift>5</DefaultShift>
            <Shift Byte="42">1</Shift>
            <Shift Byte="49">2</Shift>
            <Shift Byte="52">3</Shift>
            <Shift Byte="47">4</Shift>
            <RightFragment Position="1" MinOffset="3" MaxOffset="3">02</RightFragment>
        </SubSequence>
    </ByteSequence>
</InternalSignature>
```

.

Project Dependences

The successful men of action are not sufficiently self-observant to know exactly on what their success depends.

Joseph Jacobs (1854-1916)

JaNEME's Additional Maven Dependencies

Maven Artifact	Maven GroupId	Version
netcdf-java-all	essi-unidata	4.1
velocity	org.apache.velocity	1.6.2
velocity-tools	org.apache.velocity	2.0-alpha1
derby	org.apache.derby	10.5.3.0_1
mail	javax.mail	1.4.1
spring-jdbc	org.springframework	2.5.3
spring-aop	org.springframework	2.5.3
cglib	cglib	2.2
asm	asm	3.2

JaNeME's Functionality

Diversity is the one true thing we all have in common. Celebrate it every day.

JaNEME's Functionality

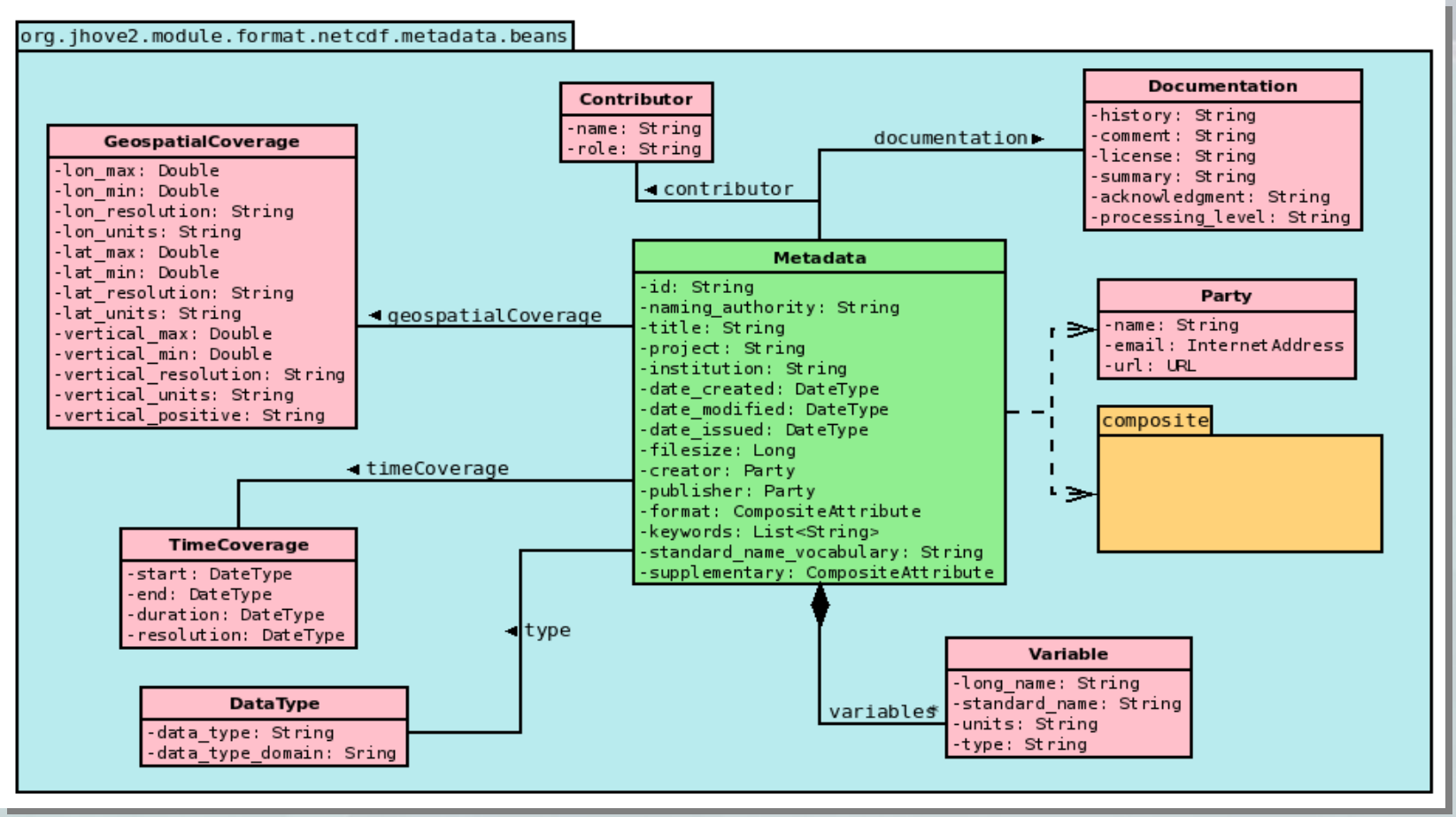
1) Platform independent

JaNeME runs on different operating systems without incurring any complicated modifications.

It has been tested on Windows Vista, Ubuntu 9.x, and Solaris 5.10.

JaNEME's Functionality

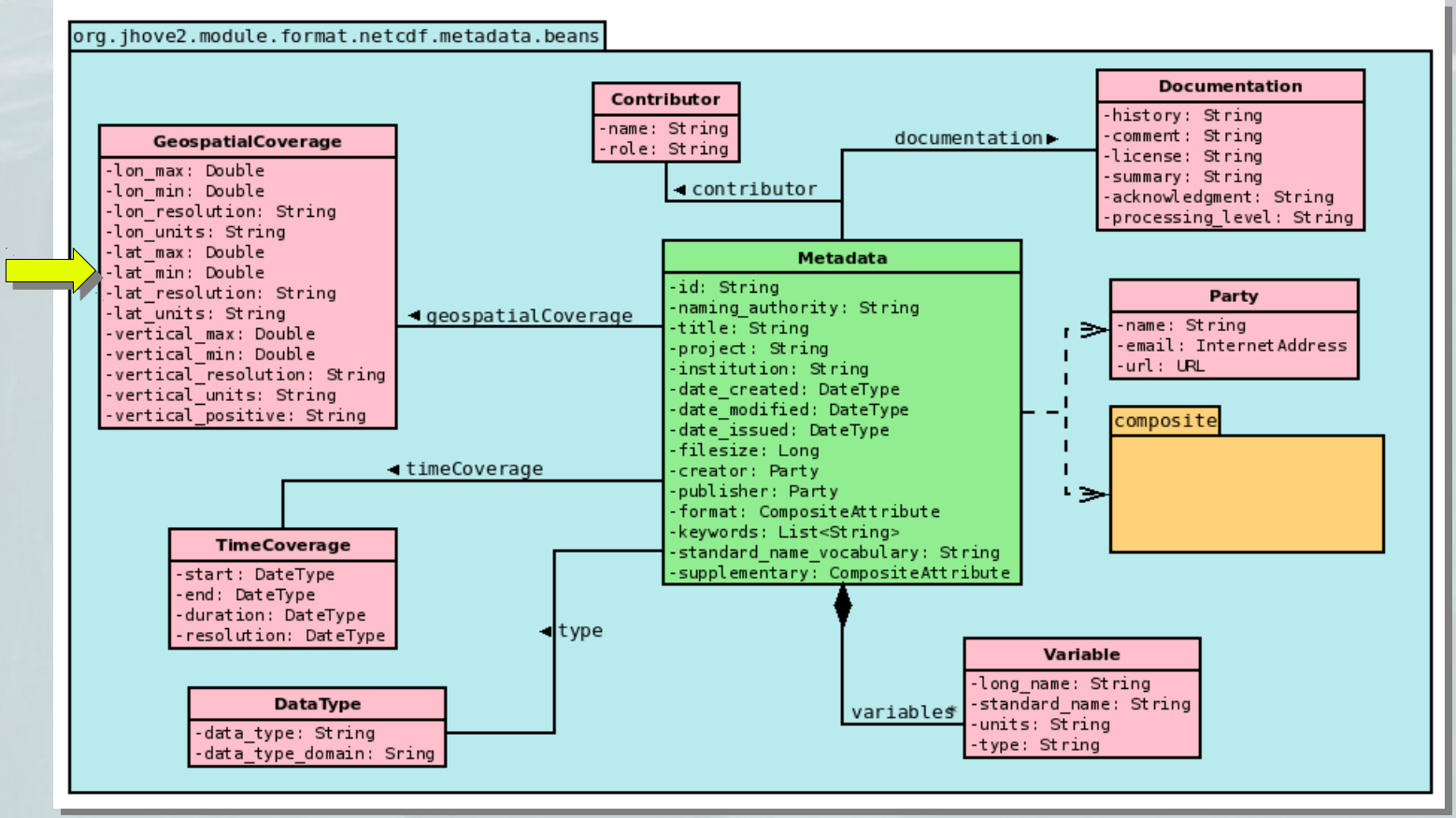
2) Header Attribute Mapping based on the Unidata Dataset Discovery v1.0 Convention [con][map]



Class Diagram for Metadata Beans

JaNEME's Functionality

2) Header Attribute Mapping based on the Unidata Dataset Discovery v1.0 Convention [con][map]



Example: Locate the *geospatial_lat_min* attribute mapping.

netcdf.metadata.*geospatial_lat_min*=*geospatialCoverage.lat_min*

JaNEME's Functionality

3) Customizable mappings

netcdf.metadata.[global_attribute_name]=[setter_expression]

Example:

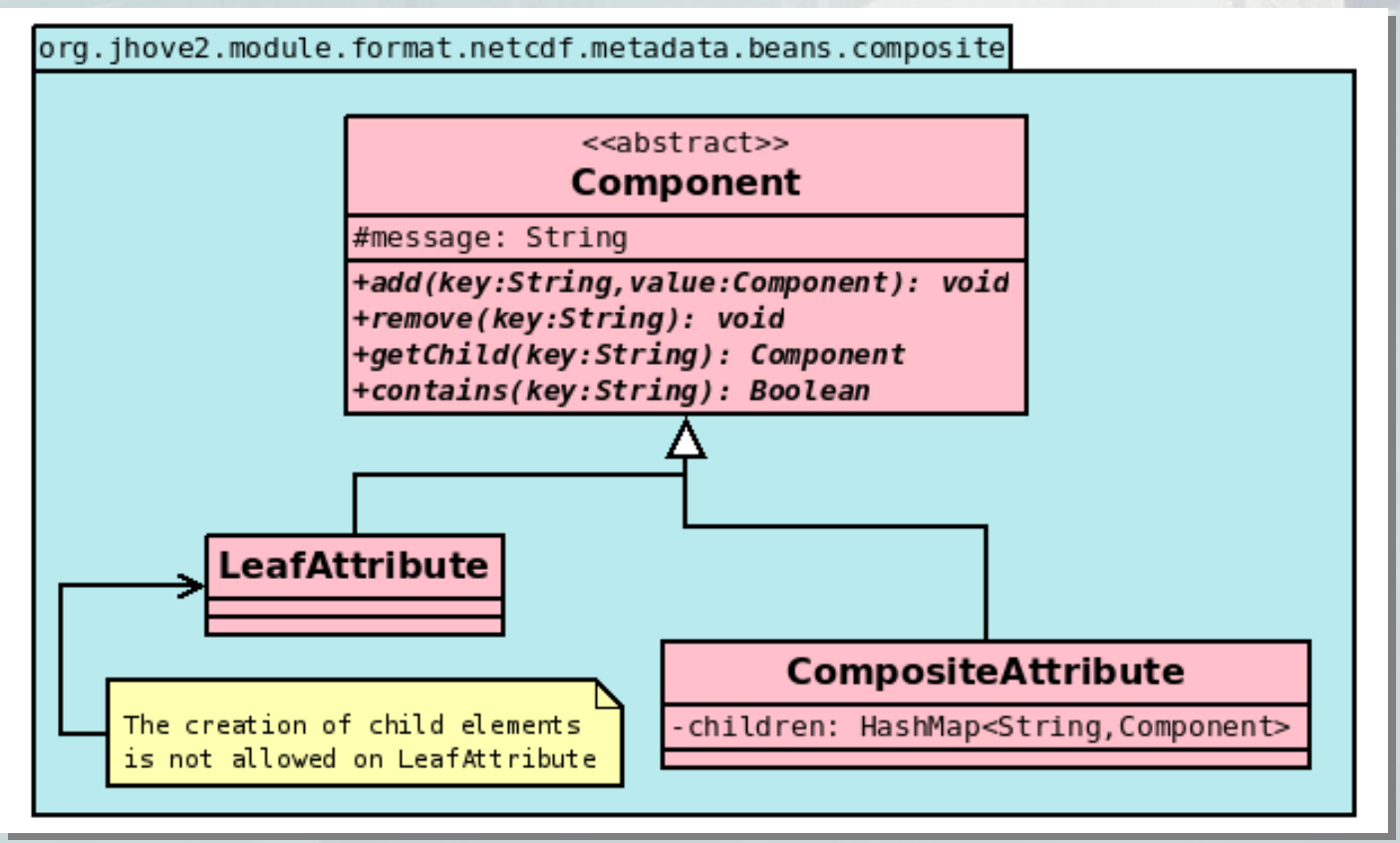
```
#mapping control /mapping steuern
netcdf.metadata.id=id
netcdf.metadata.naming_authority=naming_authority
netcdf.metadata.title=title
netcdf.metadata.keywords=keywords
netcdf.metadata.keywords_vocabulary=keywords_vocabulary
netcdf.metadata.comment=documentation.comment
netcdf.metadata.creator_name=creator.name
netcdf.metadata.creator_url=creator.url
netcdf.metadata.creator_email=creator.email
netcdf.metadata.project=project
```

Attribute name

Setter-Expression

JaNEME's Functionality

3) Customizable mappings: Composite Design Pattern



JaNEME's Functionality

3) Customizable mappings

In jhove2.sh:

JAVA_OPTS="-Dnetcdf.mappings=path to the mapping property file"

Java $JAVA_OPTS -classpath $CP org.jhove2.app.JHOVE2CommandLine $ARGS

For example:

#mapping control /mapping steuern

netcdf.metadata.software_version=supplementary.software.version

#Regular Expression for Keyword Delimiters /regulärer Ausdruck für Schlüsselworttrennzeichen

netcdf.metadata.delimiters=[,\n]+

#Path to c3grid template on disk

netcdf.displayer.c3grid.template=/edvs1/user1/jmejia/workspace/jhove2/test.xml

JaNEME's Functionality

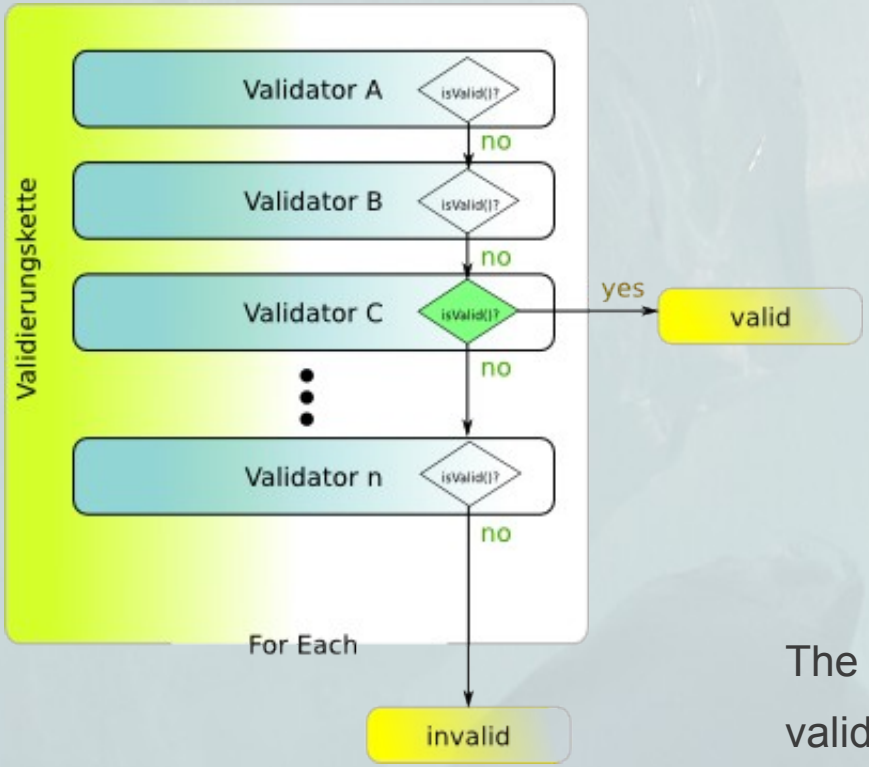
4)Integration of new vocabularies for variable and attribute names

The CFValidator Vocabulary makes use of a Apache Derby Database with 1907 Standard names und 145 Aliases conforming to the *NetCDF Climate and Forecast (CF) Metadata Convention 1.4* [cfc] .

JaNEME's Functionality

5) Validation

Through the application of the metadata bean annotations *@RequiredProperty* and *@Validate* and a Around-Advice-Interceptor (Aspect Oriented Programming)



A Validation Chain proves the conformity of variable names and attribute values to certain vocabularies.

The end user will receive a warning when the validation fails (result= invalid)

JaNEME's Functionality

6) Templating

The profiles C3grid (iso19115/139) [c3p] and Dublin Core [dc] are included.

./jhove2.sh -d c3grid -o output.xml /home/user/Beispiel.nc

or

./jhove2.sh -d dublincore -o output.xml /home/user/Beispiel.nc

JaNEME's Functionality

7) Web Service

•JaNeME includes an axis2 SOAP 1.1 web service with document binding, exposing the operations *getVersion* and *characterize*

•*getVersion* returns JHOVE2`s version

•*characterize* replies with the XML displayer output for a MTOM attached binary file argument. This file is stored in JHOVE2 temporary directory and erased after being parsed. It is important to notice the need for enabling MTOM on the client side.

JaNEME's Functionality

7) Web Service

Its deployment consists of the following steps:

a) Installation of a Servlet Container (i.e. Apache Tomcat 6.0) or axis2 Binary Distribution 1.5.1+

b) Extension of the container`s classpath to include the path of JHOVE2 config and config/droid directories on the filesystem. This usually demands the modification of some startup batch script.

c) Generate the web service .aar file in the project's target directory with the maven command: mvn org.apache.axis2:axis2-aar-maven-plugin:aar

d) If using the axis2 binary distribution, it suffices to deploy the service .aar file in $AXIS2_HOME/repository/services

e) If using a Servlet Container, deploy the .war archive distribution of axis2. Its web application will then guide you through the process of uploading the generated .aar file with the browser.

References

•[awi] Alfred Wegener Institute for Polar and Marine Research http://www.awi.de

•[c3p] C3Grid ISO 19115 Metadata Profile
http://www.c3grid.de/fileadmin/c3outreach/generation-1/metadata_profile.pdf

•[cfc] NetCDF Climate and Forecast (CF) Metadata Convention 1.4
http://cf-pcmdi.llnl.gov/documents/cf-standard-names/

•[con] NetCDF Attribute Convention for Dataset Discovery
http://www.unidata.ucar.edu/software/netcdf-java/formats/DataDiscoveryAttConvention.html

•[dc] Dublin Core http://dublincore.org

•[dro] DROID http://droid.sourceforge.net/

•[gb] GRIB on Wikipedia http://en.wikipedia.org/wiki/GRIB

•[grib-FM92-IX] A guide to the code form FM 92-IX Ext. GRIB
http://www.wmo.int/pages/prog/www/WDM/Guides/Guide-binary-2.html

•[grib-FM92-XII] Guide to the WMO Table Driven Code Form Used for the Representation and Exchange of Regularly Spaced Date in Binary Form: FM 92 GRIB Edition 2
http://www.wmo.int/pages/prog/www/WMOCodes/Guides/GRIB/GRIB2_062006.pdf

References

•[jho] JHOVE2 https://confluence.ucop.edu/display/JHOVE2Info/Home

•[jnc] NetCDF Java Library http://www.unidata.ucar.edu/software/netcdf-java/

•[lza] Langzeitarchivierung von Forschungsdaten (Wissgrid) http://www.wissgrid.de/workgroups/ap3.html

•[map] Metadata Mappings
http://www.unidata.ucar.edu/software/netcdf-java/formats/ncACDD-metadataMappings.html

•[nc] NetCDF network Common Data Form http://www.unidata.ucar.edu/software/netcdf/

•[Pronom] The technical registry Pronom http://www.nationalarchives.gov.uk/PRONOM/Default.aspx