Noumea : a Model-Driven Framework for NetCDF-CF Data Extraction and Analysis

Jean-Philippe Babau, Jannaï Tokotoko, Oumar Kande,

Abdallahi Bilal

Lab-STICC / UBO / UEB, Brest, France

babau@univ-brest.fr

NetCDF, standards and tools

- **NetCDF** : a generic file standard for array-oriented scientific data
  - File format (machine independent)
  - Data (variable and dimensions) and metadata (attributes)
  - Data access libraries (Java, C, C++, Fortran, Python, …)

- Standards for environmental data (Climate and Forecast, OceanSITES, …)
  - Conventions about variables and attributes (naming, types, …)
  - Considering space and time dimensions

- Is a NetCDF standard compliant?
  - Can we correct it to respect the standard ?
- What kind of data are present in a NetCDF file ?
  - Time series, trajectories, profiles, …
- What can we do with data ?
  - Visualization, analysis tools

Outlines

- Problem Overview
- The Noumea approach
- Conclusion and Ongoing Works

The Noumea approach : 5 STEPS

- Model Driven representation
  - A concept definition is a class with constraints
- Verification
  - Standard
  - Longitude, Latitude, Vertical, Time axis
  - Correction
- Normalize
  - Add information, if necessary
  - Naming convention
  - Unity
- Optimal storage
- Patterns extraction
  - Time series, profiles, trajectories, …
- Usage based on generic reusable tools
  - CMS-like tools
  - Visualization and analysis tools

Verification

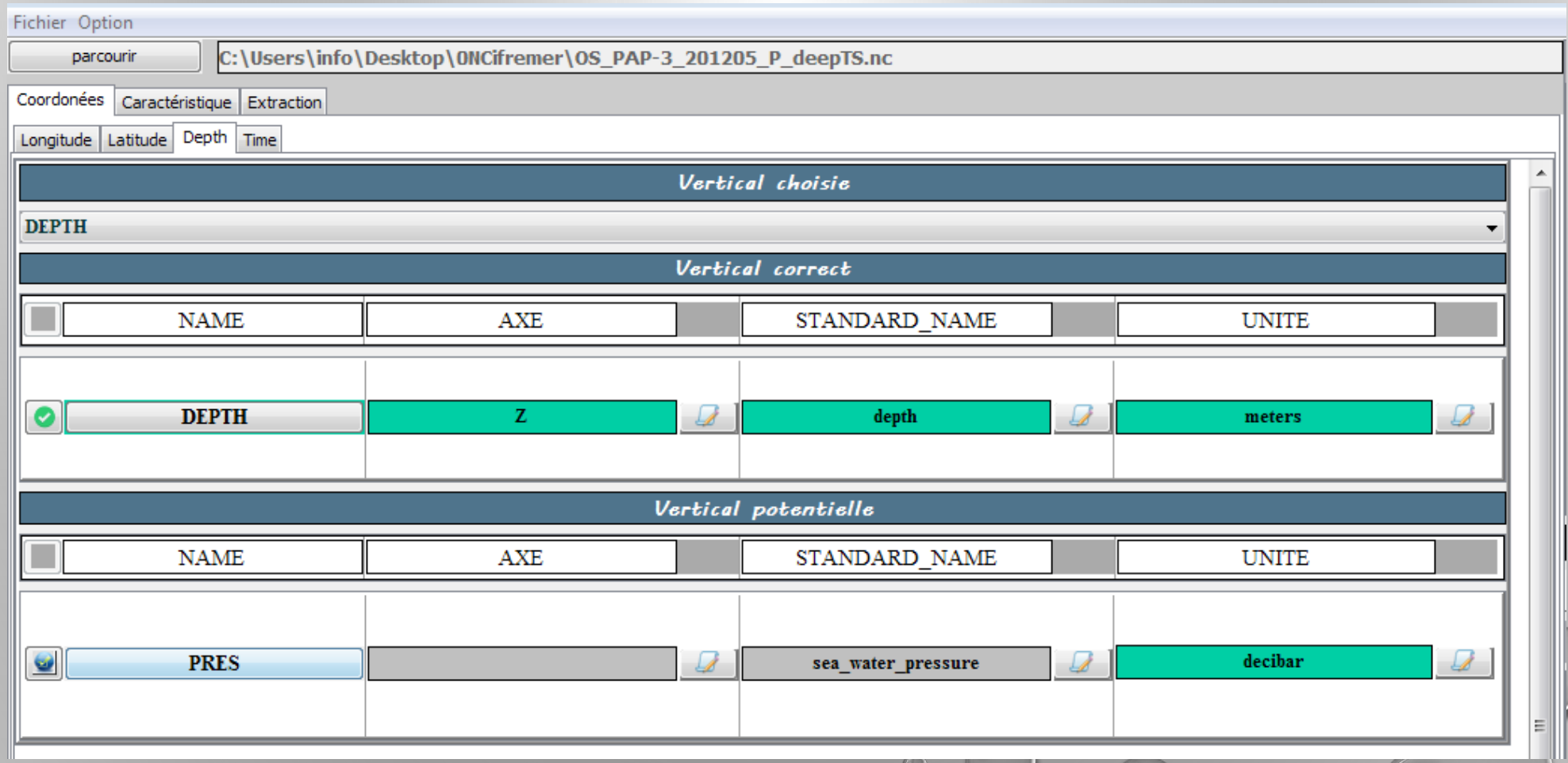
- Based on standards
- C&F example : existing coordinates and time axis
  - One variable for each axis (Longitude, Latitude, Vertical, Time)
  - One corresponding dimension
  - Coordinates metadata : unit, standard_name, axis (x, y, z, t)

- Verification principles
  - Three kinds of variable list: *Ok*, *Potential*, *not valid*
    - Ok : the three features are correct
    - Potential : at least one feature is correct
    - Not valid : no feature is correct
  - One variable selected from the Ok list
  - User interaction on Ok and Potential list
    - Status view
    - Modification of one metadata
      - Help on editing feature : proposition of correct value
  - Dynamic adaptation of Ok, Potential, Not Valid lists

Verification

- Other cases
  - Grid, contiguous sets
  - Coordinate only for reference points
- Not existing coordinate
  - Add a by-default variable (and dimension 1)
    - Brest, level 120, today
  - Editable value

A verification and correction tool



Verification

- OceanSITES example
  - Constraints on attribute
  - Naming constraints
- Verification principle
  - Constraints expressed using OCL

OceanSITES

Global Attribute

time_coverage_ end	time_coverage_end="2006-03-05T23:59:29Z"	Final date of the data in UTC.

Where time is specified as a string, the ISO8601 standard **"YYYY-MM-DDThh:mm:ssZ"**

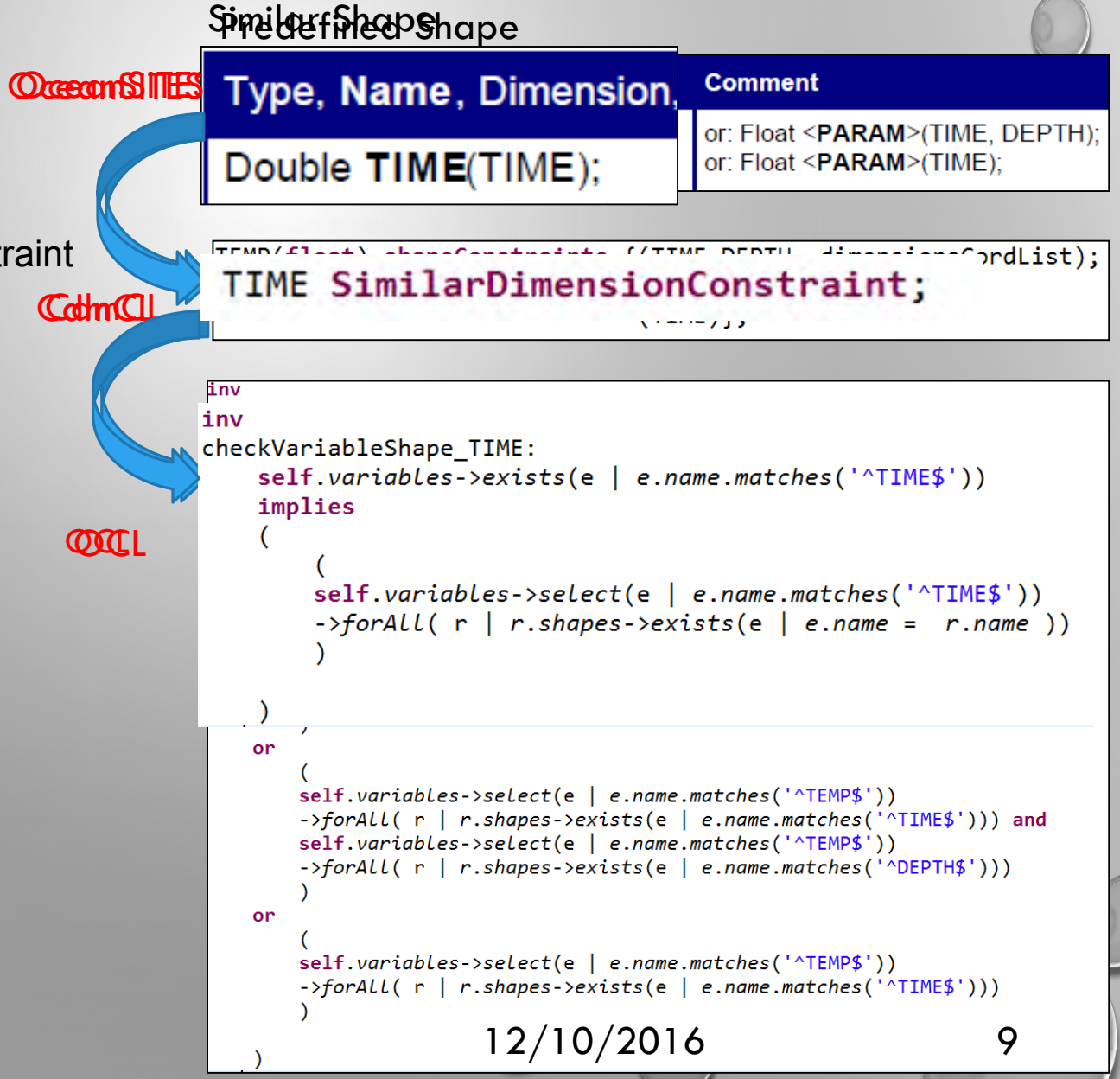
CdmCL

```
time_coverage_end value isStandardizedBy (ISO8601);
```

OCL

```
inv checkGlobalAttributeValue_time_coverage_end:
self.globalAttributes->exists(e | e.name.matches('^time_coverage_end$'))
implies
self.globalAttributes->select(e | e.name.matches('^time_coverage_end$'))->forAll(g |
g.value.matches('^\\d{4}-[0-1][0-3]-[0-3]\\d{1}T[0-2]\\d{1}:[0-5]\\d{1}:[0-5]\\d{1}Z$')
)
```

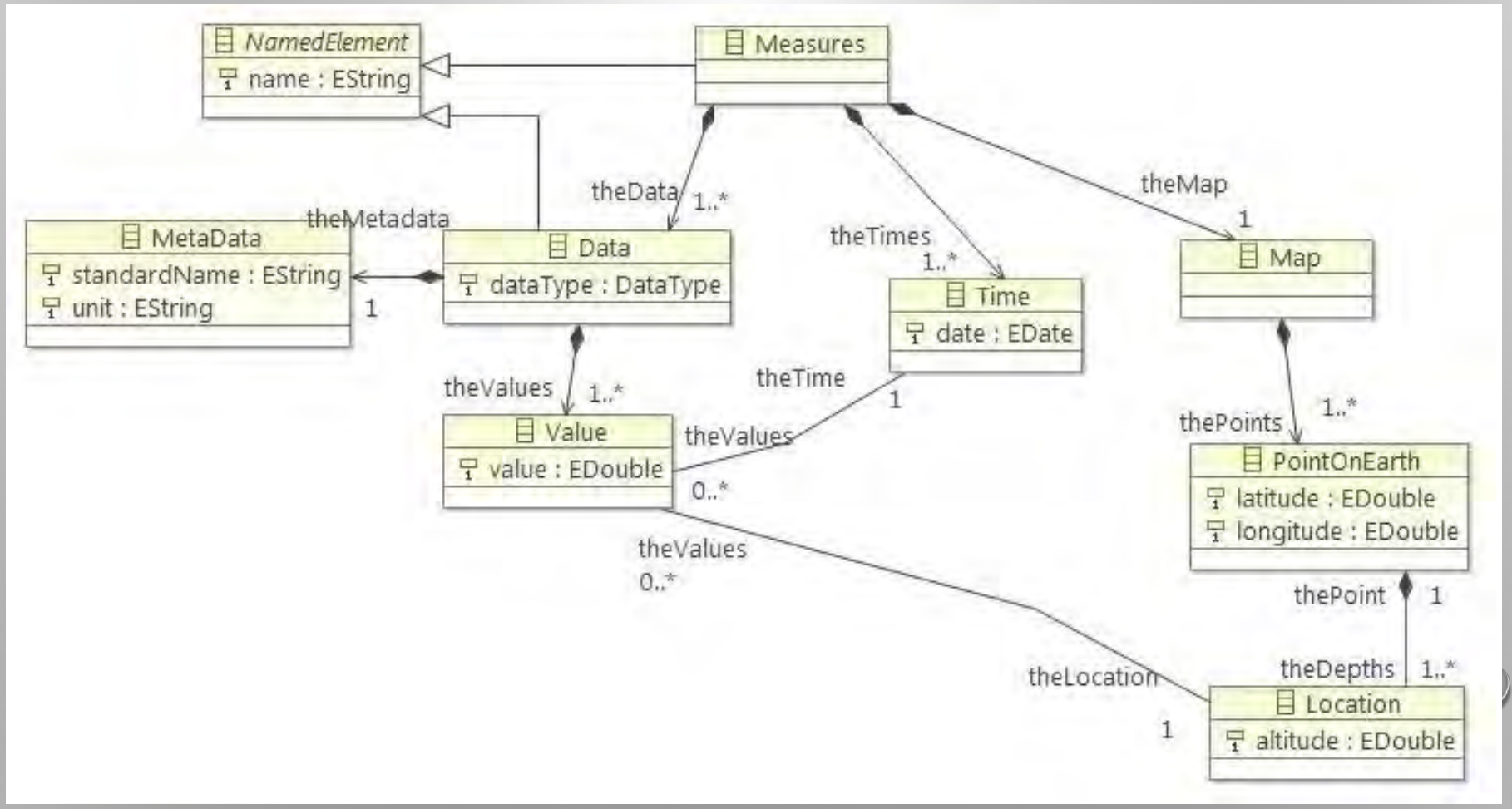
- Variable shape
  - SimilarDimensionConstraint
  - PredefinedShape



Normalization

- After the verification step
- Coordinates and time axis
  - name (not in CF) and standard_name
  - Unity and type
- Other variables
  - Attributes
- Grids, compression
  - Grid : create coordinates variables
  - Flatten contiguous sets

Generic data structure



~ UML profile Geographic information – Observations and measurements
~ Information system for « network of temperature sensors of South and South West Pacific region”

Storage policy

- Logical and physical structure are different
  - Logical : object-oriented generic data structure
  - Physical : OO, RDB, NoSQL, …
- Optimal storage policy
  - Performance dependent
- Work in progress
  - OO and/or Relational DB for metadata
  - NoSQL for data
  - Caching data mechanism

Patterns extraction

- Patterns
  - Profile : dimension for Z is strictly positive, dimension of 1 for X,Y and T
  - Time series : dimension for T is strictly positive, dimension of 1 for X,Y and Z
  - 3D : dimensions for X, Y, Z are strictly positive, dimension of 1for T
  - 2D : dimensions for X, Y are strictly positive, dimension of 1 for Z and T
  - 3D Movie : dimensions for X, Y, Z, T are strictly positive
  - 3D Trajectory : dimensions for X, Y, Z, T are strictly positive and equals, one value for each (x,y,z,t)

- Automatic extraction
  - 3D : a set of profiles = {a profile for each *PointOnEarth*, when *Depth* is more than 1}

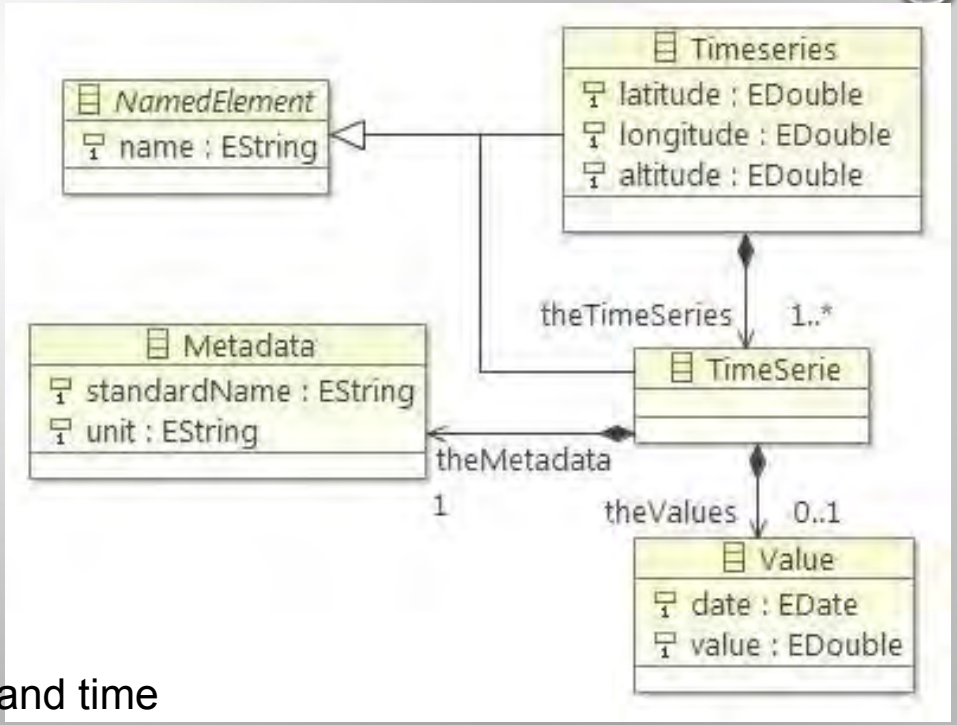
~ CSML (Climate Science Modelling Language)

Analysis tools

- Based on patterns
  - A specific input model
- Data to tools
  - Configurable : which data are used as inputs for the tool
- Generic tools
  - Configurable through a set of tool parameters

Example of time series visualization

- Input model
  - Time and variable values
- Tool parameters
  - Title, image, logo, legend
    - position
  - List of time series
    - color
  - Minimum, maximum, step for values and time
- Examples
  - Student number in Master
  - Temperature evolution



Nombre d'etudiants

Evolution des effectifs d'étudiants de Master 2 SIAM au cours des 5 dernières années

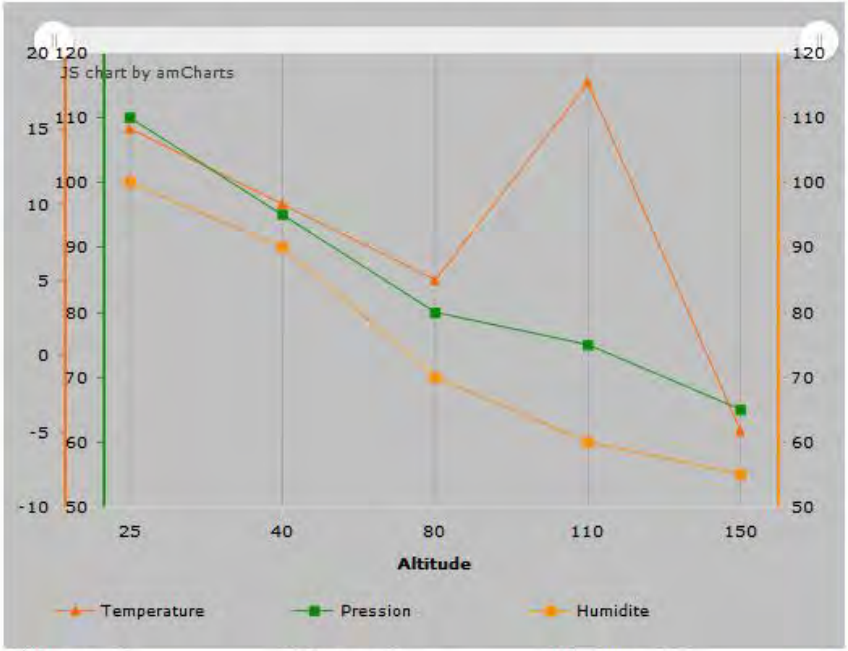
Effectif minimal : 16
Effectif maximal : 22
unité : Etudiant

Viewer Data Netcdf

IRD
Institut de recherche pour le développement

Représentation graphique de l'évolution générale des données météorologiques

Longitude :
-4.49
Latitude :
48.39
Verticale :
120.0



Temperature	Pression	Humidite
Minimum : -5	Minimum : 65	Minimum : 55
MAximum : 18	MAximum : 110	MAximum : 100
Unité : degré Celsius	Unité : Pascal	Unité : (%)

Modeling approach

- Model-Driven Engineering
  - All is model (objects in our case)
  - Meta-model (Class diagram and OCL expressions in our case)

- Not a central and unique model but… a set of concern-oriented models
  - File format model (CDM to model NetCDF)
  - Verification model (CdmCL based on OCL)
  - Logical model (georeferenced and timed measures)
  - Storage policy models (performance oriented)
  - Tool-oriented models (what is necessary for the tool? what parameters can be adapted?)

Conclusion

- NOUMEA : a global approach to deal with data
  - NetCDF verification and normalization
    - Axis and metadata constraints
  - Models at different levels
    - Object-oriented
  - Automatic pattern extraction
    - Time series, trajectories, profiles
  - Generic tooling
    - Based on patterns
    - Configurable
    - Adaptable to others data and standards

On Going Works

- Other file formats and standards (CSV, XML, …)
- Big data integration
  - NoSQL
- New generic analysis and visualization tools
  - The GeoCMS tool for visualization
- New applications
  - Quality Control and statistical analysis
- Data and tools interoperability
- A French initiative with representative actors
  - Open to international collaborations

THANKS

BIBLIOGRAPHY

- Geographic information – Observations and measurements (ISO 19156)
  - UML profile for observations
    - Generic model for scientific observations
  - Based on UML
  - 2013, OGC
- CSML (Climate Science Modelling Language)
  - OCL constraints to determine type of geometric data
    - grid, profile, trajectory, timeseries, section, …
  - Based on CF and CDM
  - 2011, OGC
- Information system for « network of temperature sensors of South and South West Pacific region”
  - A data exchange and management service
  - Generic simple data and metadata model for measures
  - 2012, GOPS (Grand Observatoire de l’environnement et de la biodiversité terrestre et marine du Pacifique Sud)