NetCDF
as a format for the output of semantic pipelines in biodiversity and ecosystem studies: AnaEE RI

Christian PICHOT et al.

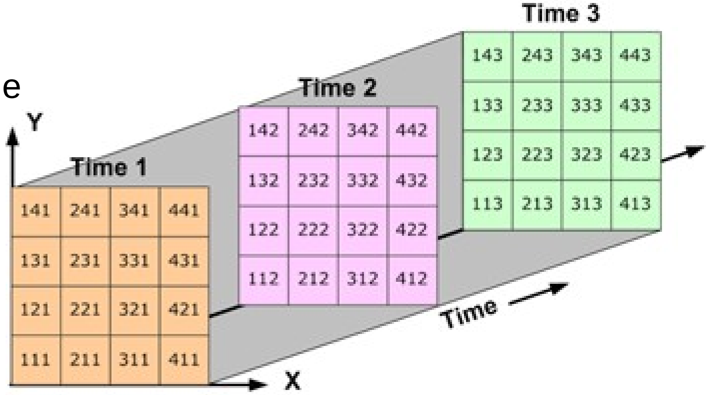
"Further Development of the DDI Cross Domain Integration Model for FAIR Data Sharing across Discipline and Domain Boundaries".
Dagstuhl Workshop, 21st September 2021

NetCDF at a glance (1/4)

- A multidimensional array of data

longitude

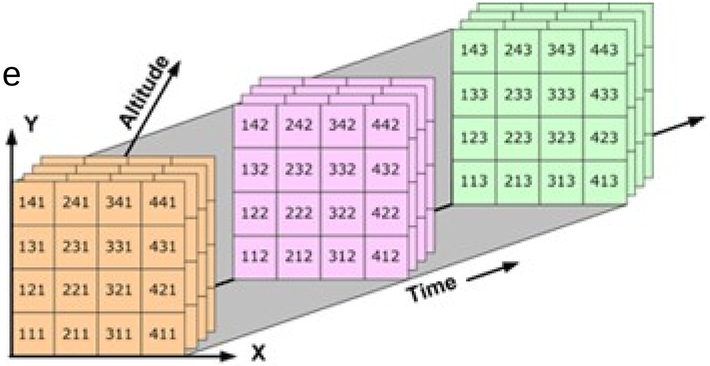
3D



latitude

longitude

4D



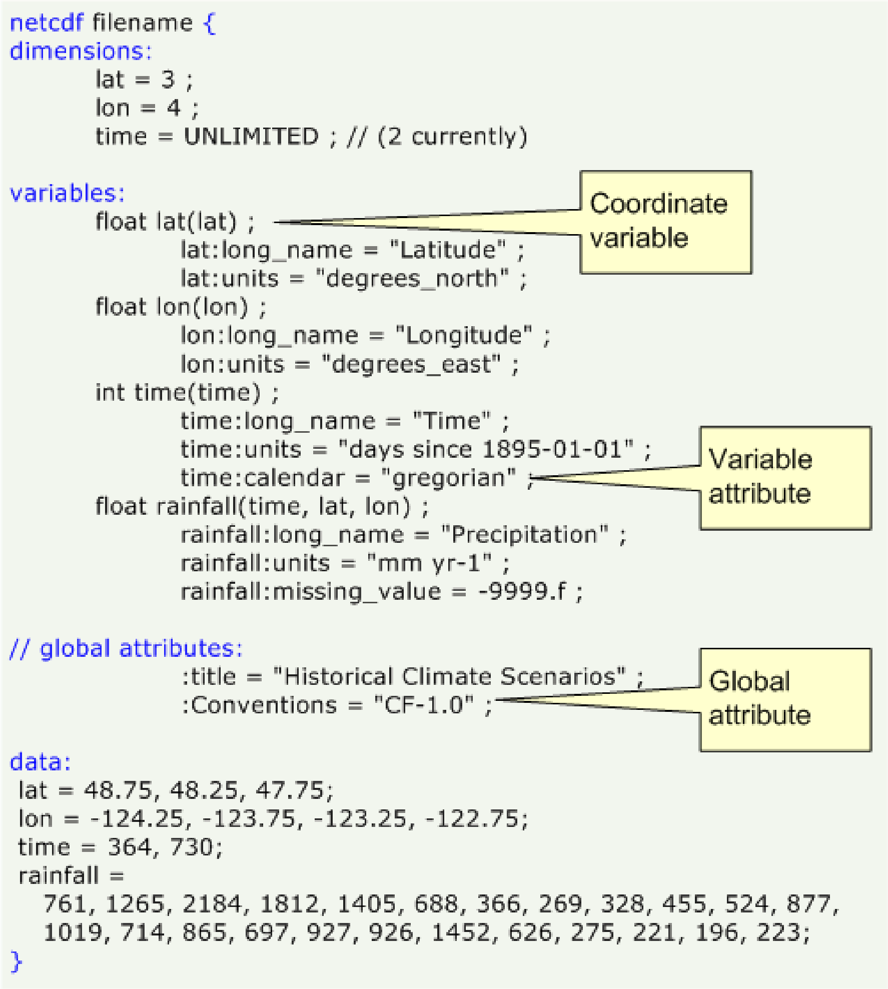
latitude

NetCDF at a glance (2/4)

- A self descriptive data file as it includes a header with dimensions, variables, and attributes

NetCDF dataset is

a header
and a data compartment



```
netcdf filename {
dimensions:
        lat = 3 ;
        lon = 4 ;
        time = UNLIMITED ; // (2 currently)

variables:
        float lat(lat) ;
                lat:long_name = "Latitude" ;
                lat:units = "degrees_north" ;
        float lon(lon) ;
                lon:long_name = "Longitude" ;
                lon:units = "degrees_east" ;
        int time(time) ;
                time:long_name = "Time" ;
                time:units = "days since 1895-01-01" ;
                time:calendar = "gregorian" ;
        float rainfall(time, lat, lon) ;
                rainfall:long_name = "Precipitation" ;
                rainfall:units = "mm yr-1" ;
                rainfall:missing_value = -9999.f ;

// global attributes:
                :title = "Historical Climate Scenarios" ;
                :Conventions = "CF-1.0" ;

data:
 lat = 48.75, 48.25, 47.75;
 lon = -124.25, -123.75, -123.25, -122.75;
 time = 364, 730;
 rainfall =
   761, 1265, 2184, 1812, 1405, 688, 366, 269, 328, 455, 524, 877,
   1019, 714, 865, 697, 927, 926, 1452, 626, 275, 221, 196, 223;
}
```

NetCDF at a glance (3/4)

- Data encoding into machine-independent sequences of bits

using the XDR (eXternal Data Representation) standard protocol.

NetCDF is

a data abstraction for array-oriented data access
AND a software library implementing the interfaces that support that abstraction.

NetCDF at a glance (4/4)

- Lightweight files
contiguous or chunked storage structure
binary
compression.

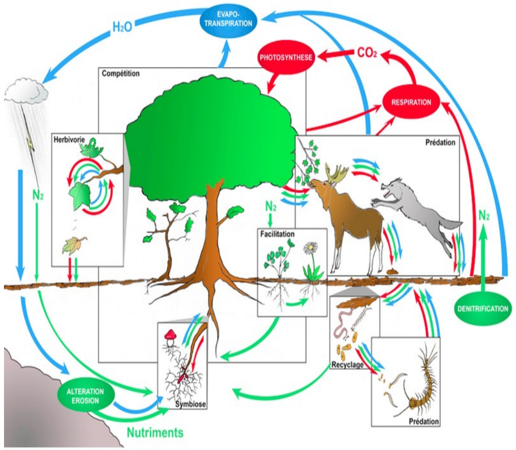
NetCDF

a self-describing, machine-independent binary data format that supports
the creation, access, and sharing of array-oriented data

Use of NetCDF

in AnaEE biodiversity and ecosystem studies

Rationale



Ecosystem study requires complex research and deals with heterogeneous, varied and widespread data.

The proper understanding and interoperability of
the information sources remains one of the greatest challenges

AnaEE
Analysis and Experimentation on Ecosystems

A Research Infrastructure for experimentation on ecosystems

ENVRI plus

ENVRI FAIR

Experimentation platforms

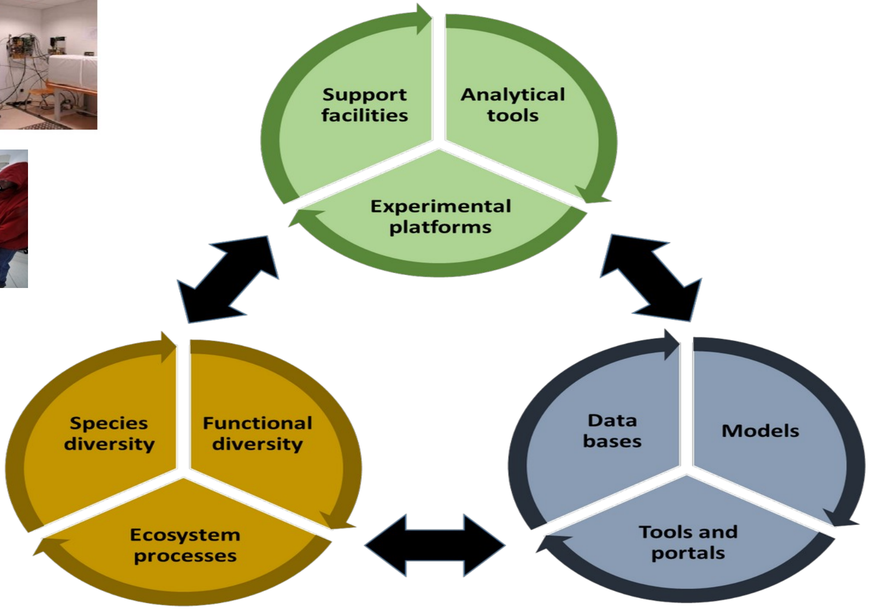


Plate-forme
de modélisation et de simulation
des agro-écosystèmes

Mobile instruments and analytical plaforms

Data and models

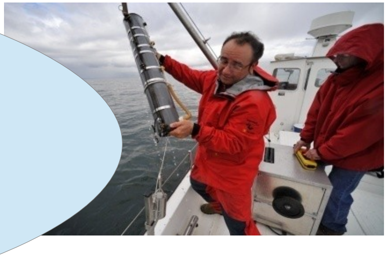
How to deal with data heterogeneity?

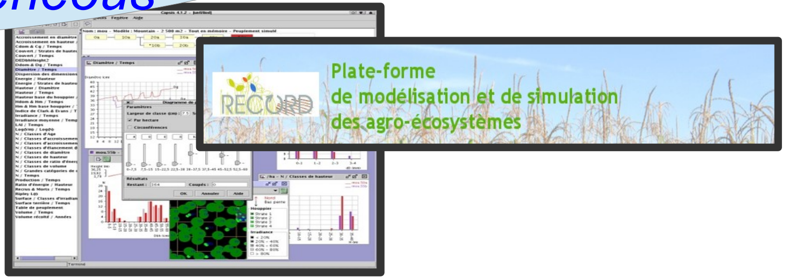
AnaEE
Analysis and Experimentation on Ecosystems
France
Laboratoire Mobile
Mesure de CO$_2$, CH$_4$, N$_2$O et H$_2$O

Managing data for:

- ☛ discovery
- ☛ access to resources

...distributed and heterogeneous





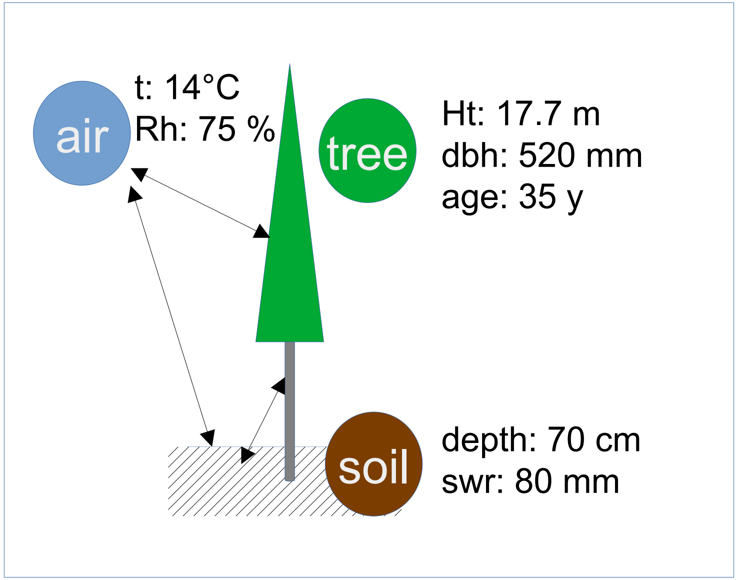
RECORD

Plate-forme de modélisation et de simulation des agro-écosystèmes

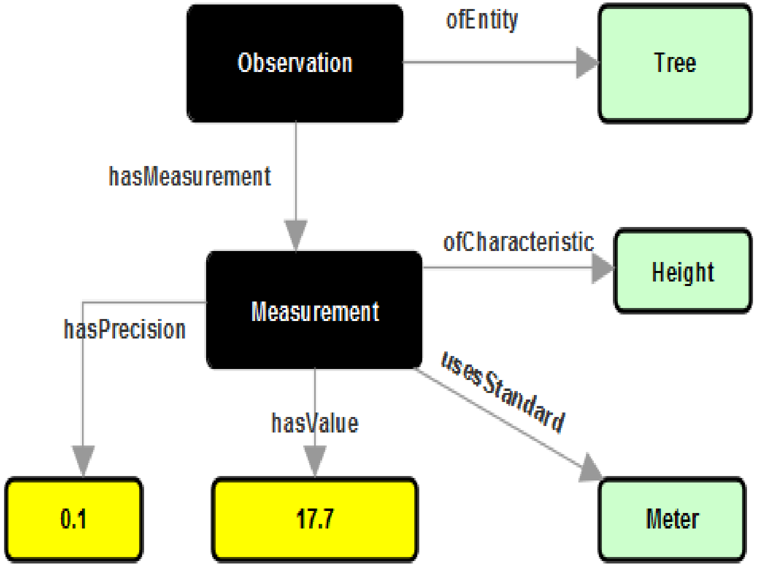
Developing semantic interoperability

Method

1) Identify
- the components of the system
- and their relationships



2) Model the system
using semantic vocabularies

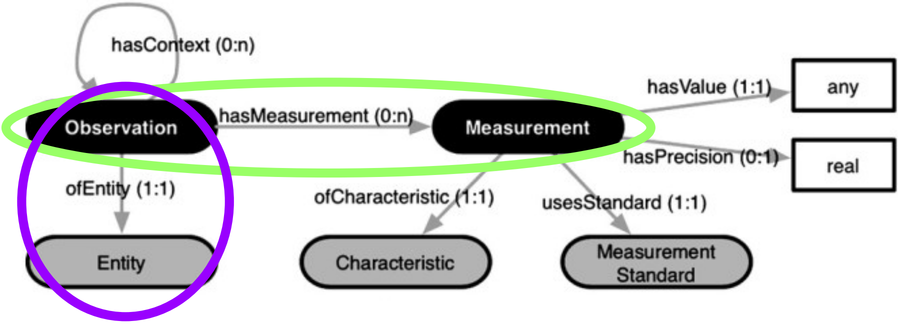


Developing semantic interoperability

Implementation

AnaEE* RI as scientific context:
The Research Infrastructure offers services for experimentation on continental ecosystems

OBOE* as ontological framework:
The ontology provides the atomic elements for modeling observations



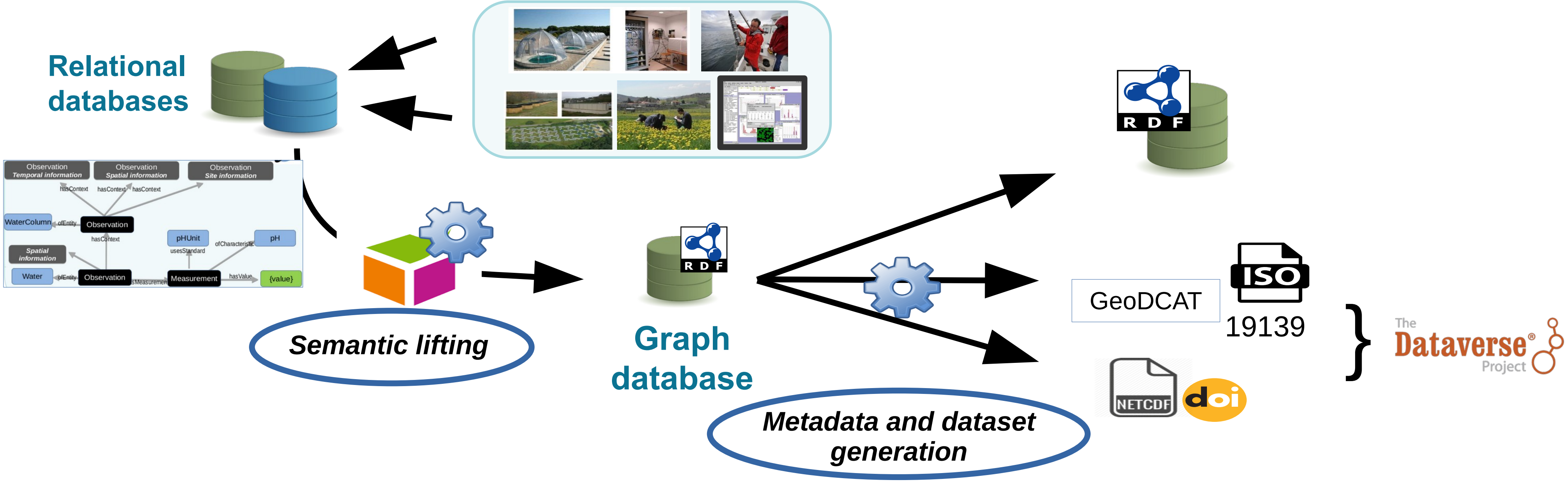
*Mark Schildhauer, Matthew B. Jones, Shawn Bowers, Joshua Madin, Sergeui Krivov, Deana Pennington, Ferdinando Villa, Benjamin Leinfelder, Christopher Jones, and Margaret O'Brien. 2016. OBOE: the Extensible Observation Ontology, version 1.2. KNB Data Repository. doi:10.5063/F1125R0F

Developing semantic interoperability

Bozen, 2021

Semantic lifting and data exploitation

Graph patterns and variable semantic descriptions are processed by a pipeline for semantic lifting of the data before their exploitation



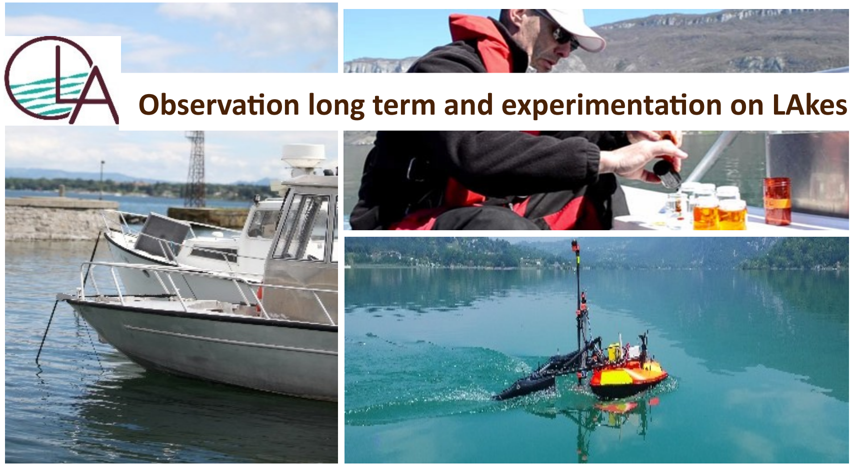
Relational databases

Semantic lifting

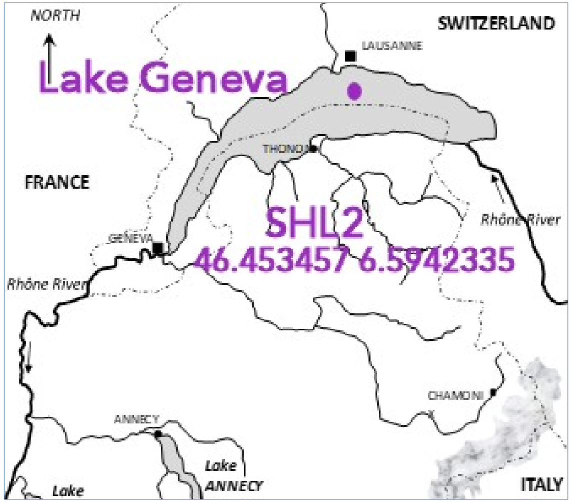
Graph database

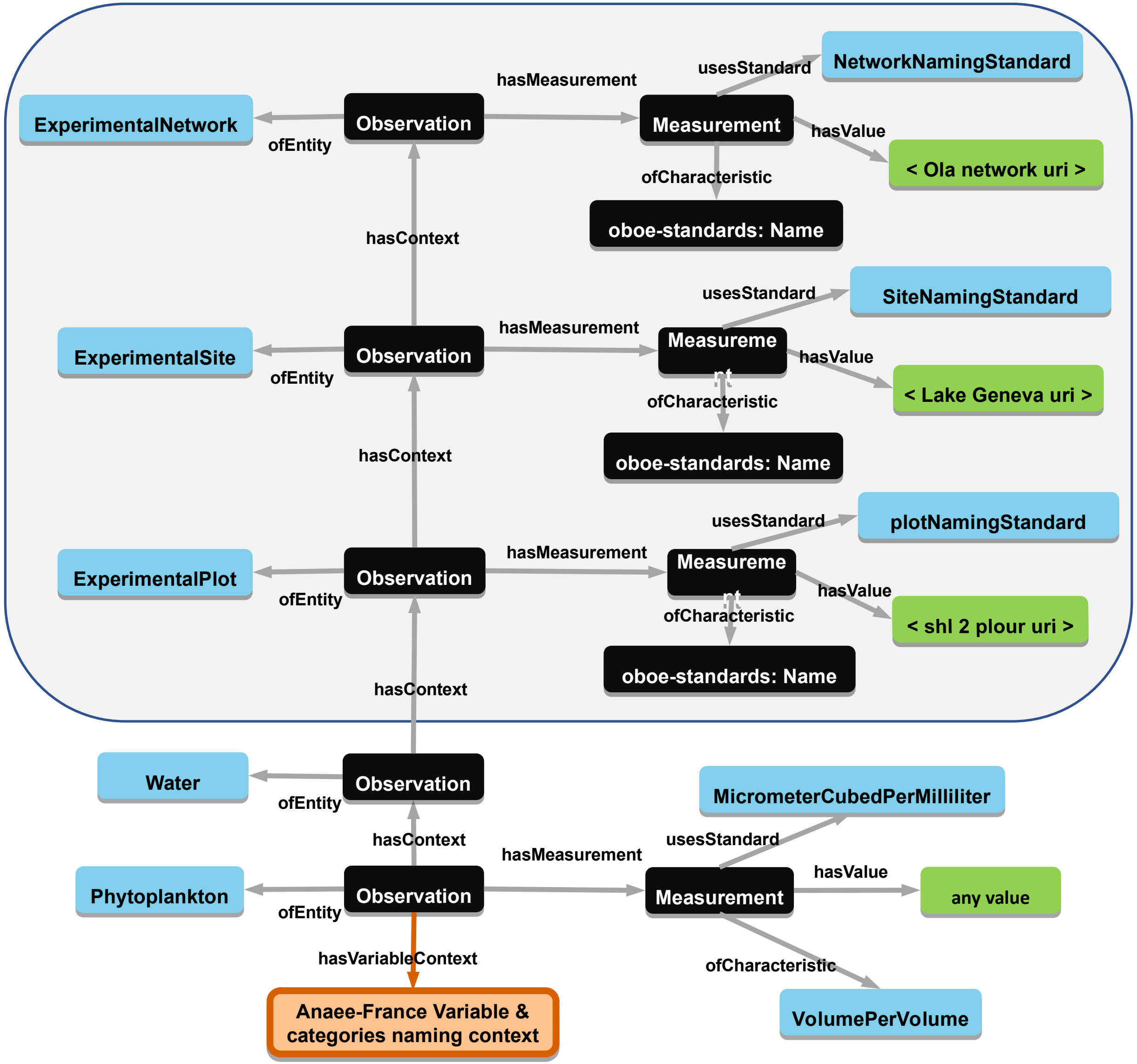
Metadata and dataset generation

Graph for phytoplankton measurements

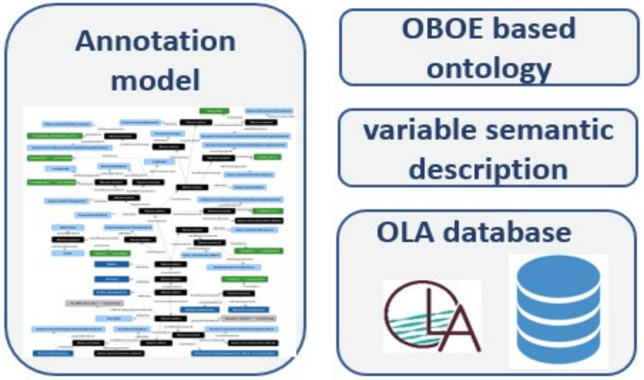


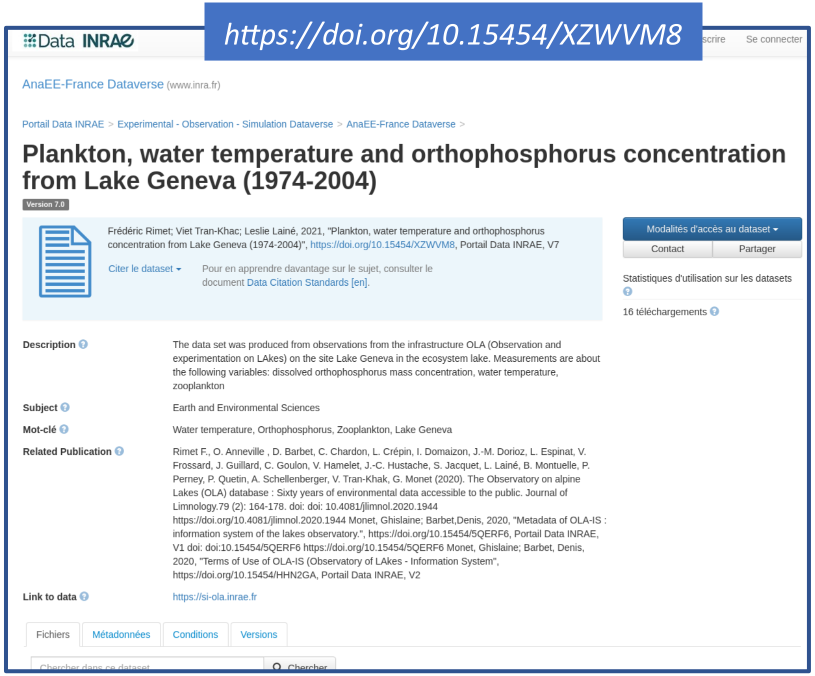
Observation long term and experimentation on LAkes





Application for planktonic biodiversity data from lakes




https://doi.org/10.15454/XZWVM8

Data INRAE — scrire — Se connecter

AnaEE-France Dataverse (www.inra.fr)

Portail Data INRAE > Experimental - Observation - Simulation Dataverse > AnaEE-France Dataverse >

Plankton, water temperature and orthophosphorus concentration from Lake Geneva (1974-2004)

Version 7.0

Frédéric Rimet; Viet Tran-Khac; Leslie Lainé, 2021, "Plankton, water temperature and orthophosphorus concentration from Lake Geneva (1974-2004)", https://doi.org/10.15454/XZWVM8, Portail Data INRAE, V7

Citer le dataset ▾ — Pour en apprendre davantage sur le sujet, consulter le document Data Citation Standards [en].

Modalités d'accès au dataset ▾ — Contact — Partager

Statistiques d'utilisation sur les datasets

16 téléchargements

Description The data set was produced from observations from the infrastructure OLA (Observation and experimentation on LAkes) on the site Lake Geneva in the ecosystem lake. Measurements are about the following variables: dissolved orthophosphorus mass concentration, water temperature, zooplankton

Subject Earth and Environmental Sciences

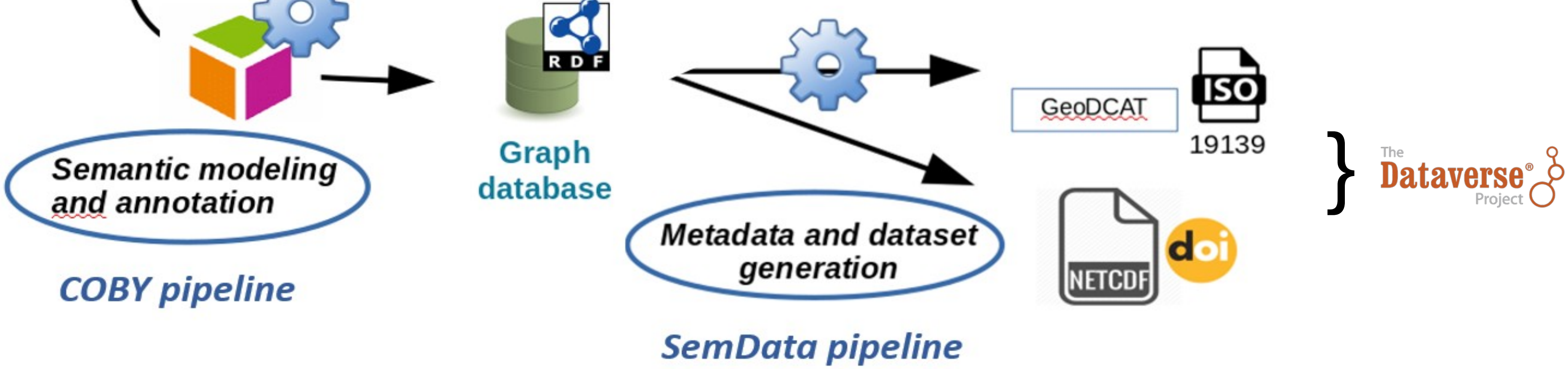
Mot-clé Water temperature, Orthophosphorus, Zooplankton, Lake Geneva

Related Publication Rimet F., O. Anneville , D. Barbet, C. Chardon, L. Crépin, I. Domaizon, J.-M. Dorioz, L. Espinat, V. Frossard, J. Guillard, C. Goulon, V. Hamelet, J.-C. Hustache, S. Jacquet, L. Lainé, B. Montuelle, P. Perney, P. Quetin, A. Schellenberger, V. Tran-Khak, G. Monet (2020). The Observatory on alpine Lakes (OLA) database : Sixty years of environmental data accessible to the public. Journal of Limnology.79 (2): 164-178. doi: doi: 10.4081/jlimnol.2020.1944 https://doi.org/10.4081/jlimnol.2020.1944 Monet, Ghislaine; Barbet,Denis, 2020, "Metadata of OLA-IS : information system of the lakes observatory.", https://doi.org/10.15454/5QERF6, Portail Data INRAE, V1 doi: doi:10.15454/5QERF6 https://doi.org/10.15454/5QERF6 Monet, Ghislaine; Barbet, Denis, 2020, "Terms of Use of OLA-IS (Observatory of LAkes - Information System", https://doi.org/10.15454/HHN2GA, Portail Data INRAE, V2

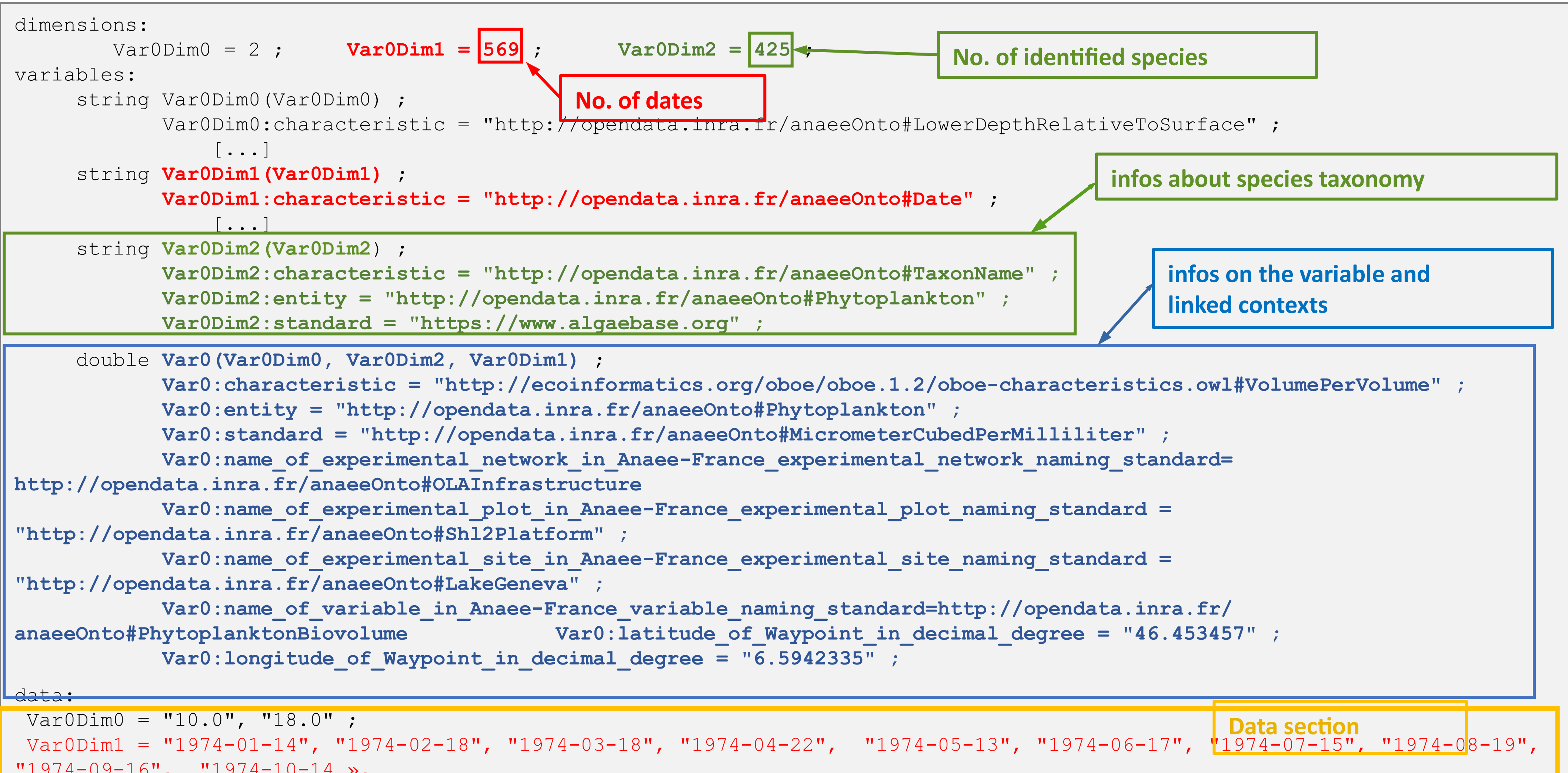
Link to data https://si-ola.inrae.fr

Fichiers — Métadonnées — Conditions — Versions

Chercher dans ce dataset — Chercher



Extract from the NetCDF file for phytoplankton biovolume (‘Var0’) expressed in MicrometerCubedPerMilliliter and provides data collected in one experimental plot (‘Shl2Platform’, 46.45°N, 6.59°E) between 10-18 depths (Dim0) expressed in meter and for 569 dates (Dim1). Phytoplankton species use the algaebase taxonomy.



```
dimensions:
        Var0Dim0 = 2 ;      Var0Dim1 = 569 ;      Var0Dim2 = 425 ;
variables:
     string Var0Dim0(Var0Dim0) ;
            Var0Dim0:characteristic = "http://opendata.inra.fr/anaeeOnto#LowerDepthRelativeToSurface" ;
                [...]
     string Var0Dim1(Var0Dim1) ;
            Var0Dim1:characteristic = "http://opendata.inra.fr/anaeeOnto#Date" ;
                [...]
     string Var0Dim2(Var0Dim2) ;
            Var0Dim2:characteristic = "http://opendata.inra.fr/anaeeOnto#TaxonName" ;
            Var0Dim2:entity = "http://opendata.inra.fr/anaeeOnto#Phytoplankton" ;
            Var0Dim2:standard = "https://www.algaebase.org" ;
     double Var0(Var0Dim0, Var0Dim2, Var0Dim1) ;
            Var0:characteristic = "http://ecoinformatics.org/oboe/oboe.1.2/oboe-characteristics.owl#VolumePerVolume" ;
            Var0:entity = "http://opendata.inra.fr/anaeeOnto#Phytoplankton" ;
            Var0:standard = "http://opendata.inra.fr/anaeeOnto#MicrometerCubedPerMilliliter" ;
            Var0:name_of_experimental_network_in_Anaee-France_experimental_network_naming_standard=
http://opendata.inra.fr/anaeeOnto#OLAInfrastructure
            Var0:name_of_experimental_plot_in_Anaee-France_experimental_plot_naming_standard =
"http://opendata.inra.fr/anaeeOnto#Shl2Platform" ;
            Var0:name_of_experimental_site_in_Anaee-France_experimental_site_naming_standard =
"http://opendata.inra.fr/anaeeOnto#LakeGeneva" ;
            Var0:name_of_variable_in_Anaee-France_variable_naming_standard=http://opendata.inra.fr/
anaeeOnto#PhytoplanktonBiovolume            Var0:latitude_of_Waypoint_in_decimal_degree = "46.453457" ;
            Var0:longitude_of_Waypoint_in_decimal_degree = "6.5942335" ;
data:
 Var0Dim0 = "10.0", "18.0" ;
 Var0Dim1 = "1974-01-14", "1974-02-18", "1974-03-18", "1974-04-22",  "1974-05-13", "1974-06-17", "1974-07-15", "1974-08-19",
"1974-09-16",  "1974-10-14 ».
```

No. of identified species

No. of dates

infos about species taxonomy

infos on the variable and linked contexts

Data section

infos about species taxonomy

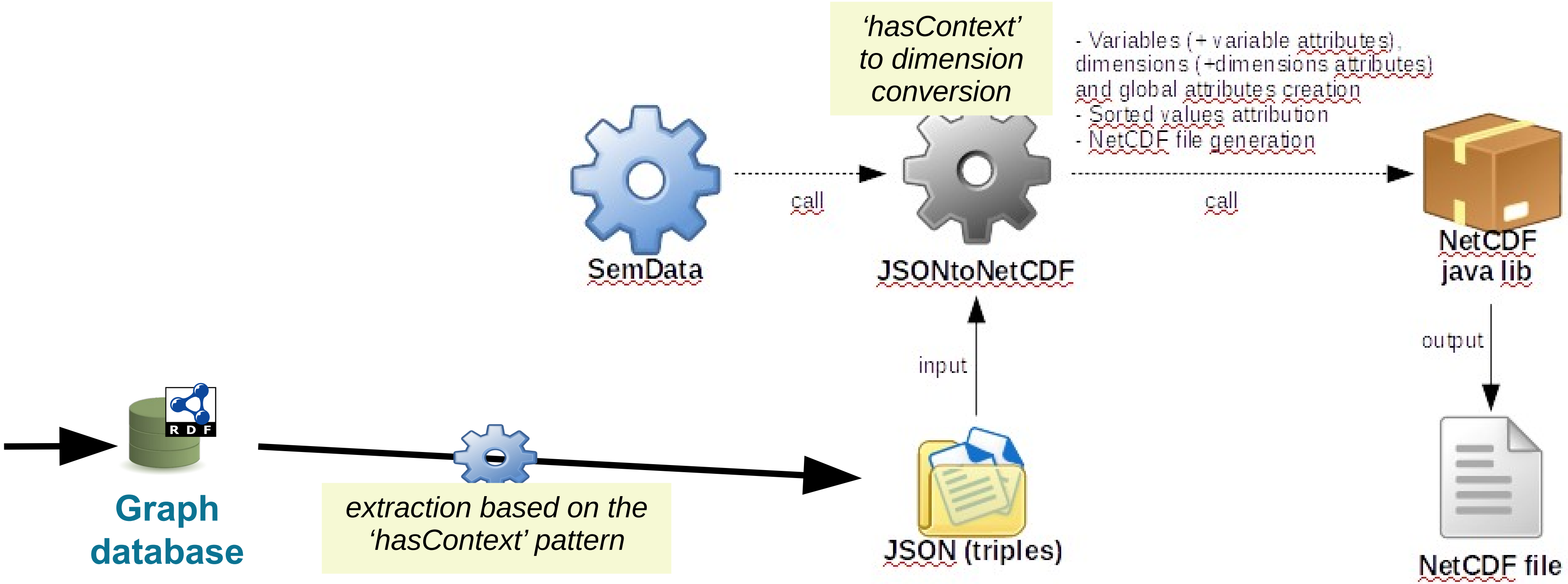
infos on the variable and linked contexts

Data section

```
                [...]
      string Var0Dim1(Var0Dim1) ;
             Var0Dim1:characteristic = "http://opendata.inra.fr/anaeeOnto#Date" ;
                [...]
      string Var0Dim2(Var0Dim2) ;
             Var0Dim2:characteristic = "http://opendata.inra.fr/anaeeOnto#TaxonName" ;
             Var0Dim2:entity = "http://opendata.inra.fr/anaeeOnto#Phytoplankton" ;
             Var0Dim2:standard = "https://www.algaebase.org" ;
      double Var0(Var0Dim0, Var0Dim2, Var0Dim1) ;
             Var0:characteristic = "http://ecoinformatics.org/oboe/oboe.1.2/oboe-characteristics.owl#VolumePerVolume" ;
             Var0:entity = "http://opendata.inra.fr/anaeeOnto#Phytoplankton" ;
             Var0:standard = "http://opendata.inra.fr/anaeeOnto#MicrometerCubedPerMilliliter" ;
             Var0:name_of_experimental_network_in_Anaee-France_experimental_network_naming_standard=
http://opendata.inra.fr/anaeeOnto#OLAInfrastructure
             Var0:name_of_experimental_plot_in_Anaee-France_experimental_plot_naming_standard =
"http://opendata.inra.fr/anaeeOnto#Shl2Platform" ;
             Var0:name_of_experimental_site_in_Anaee-France_experimental_site_naming_standard =
"http://opendata.inra.fr/anaeeOnto#LakeGeneva" ;
             Var0:name_of_variable_in_Anaee-France_variable_naming_standard=http://opendata.inra.fr/
anaeeOnto#PhytoplanktonBiovolume            Var0:latitude_of_Waypoint_in_decimal_degree = "46.453457" ;
             Var0:longitude_of_Waypoint_in_decimal_degree = "6.5942335" ;

data:
 Var0Dim0 = "10.0", "18.0" ;
 Var0Dim1 = "1974-01-14", "1974-02-18", "1974-03-18", "1974-04-22",  "1974-05-13", "1974-06-17", "1974-07-15", "1974-08-19",
"1974-09-16",  "1974-10-14 »,
 "1974-11-18", "1974-12-09", "1975-02-17", "1975-03-17",
 [...]
 Var0Dim2 = "Achnanthes catenata", "Achnanthes conspicua",  "Achnanthes exilis", "Achnanthes flexella", "Achnanthes
minutissima", "Achnanthes sp.", "Achroonema articulatum", "Actinastrum hantzschii", "Amphidinium sp.", "Amphipleura pellucida"
"Amphora ovalis",  "Amphora pediculus", "Amphora sp.
 [...]
Var0 =  NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, NaN, 399969, 222499,
328451, 603926, 111200, 31800, 74200, 0, 0, 10600, 0, NaN, 26500,
[...]
```

From rdf triples to NetCDF dataset



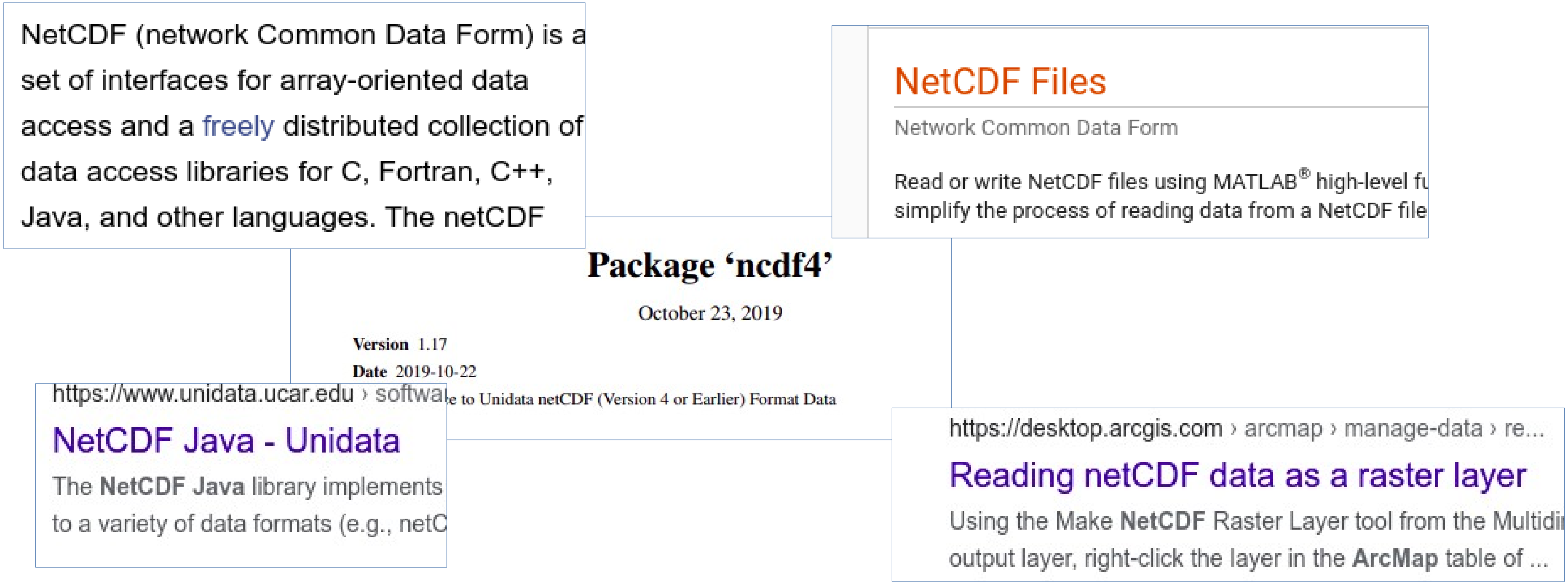
‘hasContext’ to dimension conversion

- Variables (+ variable attributes), dimensions (+dimensions attributes) and global attributes creation
- Sorted values attribution
- NetCDF file generation

extraction based on the ‘hasContext’ pattern

Reading NetCDF dataset

NetDCF library available for most programming languages, GIS and statistical environment



NetCDF (network Common Data Form) is a set of interfaces for array-oriented data access and a freely distributed collection of data access libraries for C, Fortran, C++, Java, and other languages. The netCDF

NetCDF Files

Network Common Data Form

Read or write NetCDF files using MATLAB® high-level fu
simplify the process of reading data from a NetCDF file

Package 'ncdf4'

October 23, 2019

Version 1.17

Date 2019-10-22

…e to Unidata netCDF (Version 4 or Earlier) Format Data

https://www.unidata.ucar.edu › softwa

NetCDF Java - Unidata

The **NetCDF Java** library implements
to a variety of data formats (e.g., netC

https://desktop.arcgis.com › arcmap › manage-data › re...

Reading netCDF data as a raster layer

Using the Make **NetCDF** Raster Layer tool from the Multidi
output layer, right-click the layer in the **ArcMap** table of ...

NetCDF features

https://www.unidata.ucar.edu/software/netcdf/

Data type

n-dimensional rectangular structure containing items with the same data type

Header contains:

Dimensions
Variables
Attributes

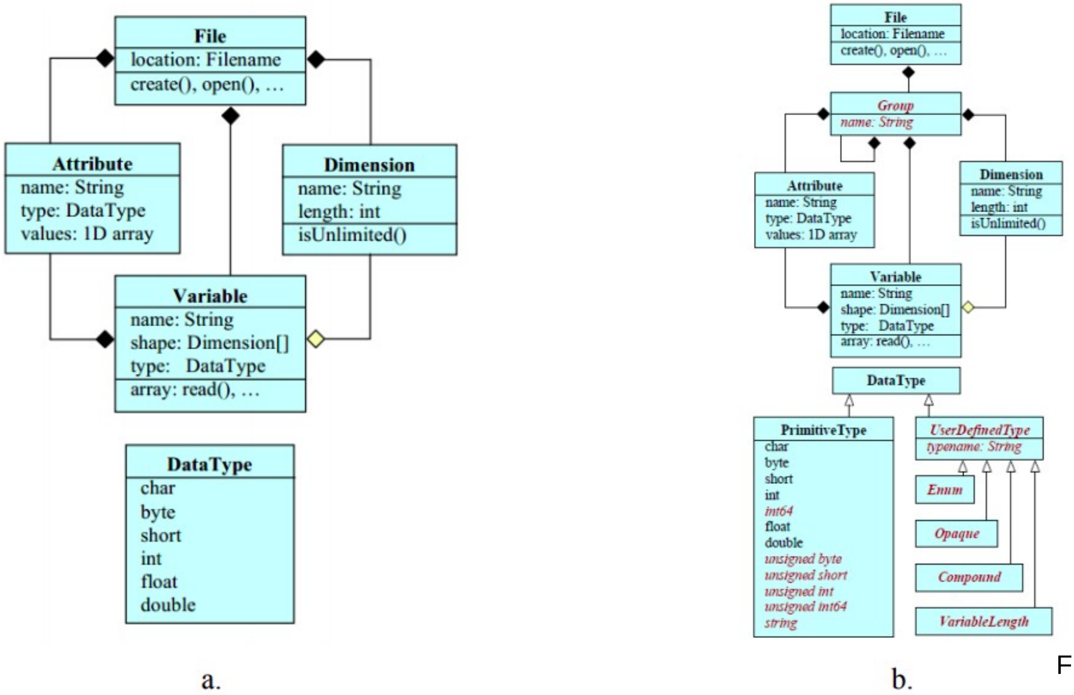
Data consists in:

a ‘fixed-sized part’ (data with limited dimension)
possibly a ‘record data part’ (data with dimenson) at the end of the NetCDF file

Additionnal features in NetCDF-4:

Group:
Compound types:
Multiple unlimited dimensions:
Chunked storage structure (HDF5)

NetCDF data models



From Haicheng Liu, 2014

Figure 2.1. a. Classic data model; b. Enhanced model with red words showing differences from classic data model (Rew et al., 2006)

https://docs.unidata.ucar.edu/netcdf-java/current/userguide/common_data_model_overview.html

NetCDF for wich community

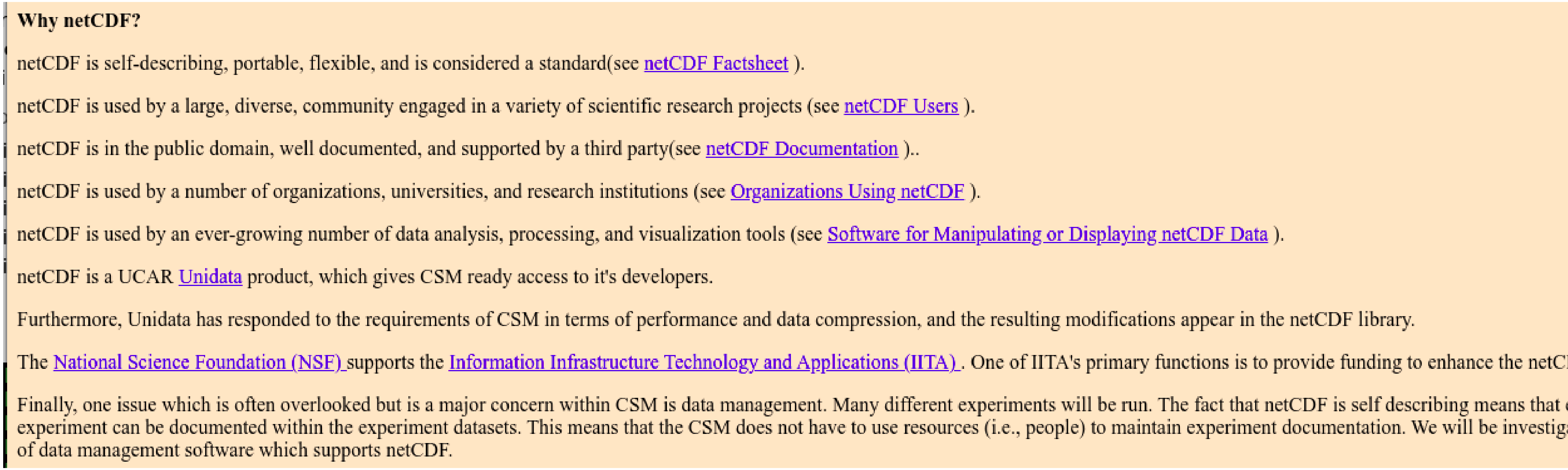
Low usage in Ecology and Biodiv as data rarely fit with a n-dimension array

Mostly used in:

- Climate community
- Ocean community

Rationale for Adopting netCDF as the Climate System Model Standard Data Format

(https://www.cgd.ucar.edu/ccr/bettge/CSM-netCDF/csm_why_netcdf.html#5)



Why netCDF?

netCDF is self-describing, portable, flexible, and is considered a standard(see netCDF Factsheet).

netCDF is used by a large, diverse, community engaged in a variety of scientific research projects (see netCDF Users).

netCDF is in the public domain, well documented, and supported by a third party(see netCDF Documentation)..

netCDF is used by a number of organizations, universities, and research institutions (see Organizations Using netCDF).

netCDF is used by an ever-growing number of data analysis, processing, and visualization tools (see Software for Manipulating or Displaying netCDF Data).

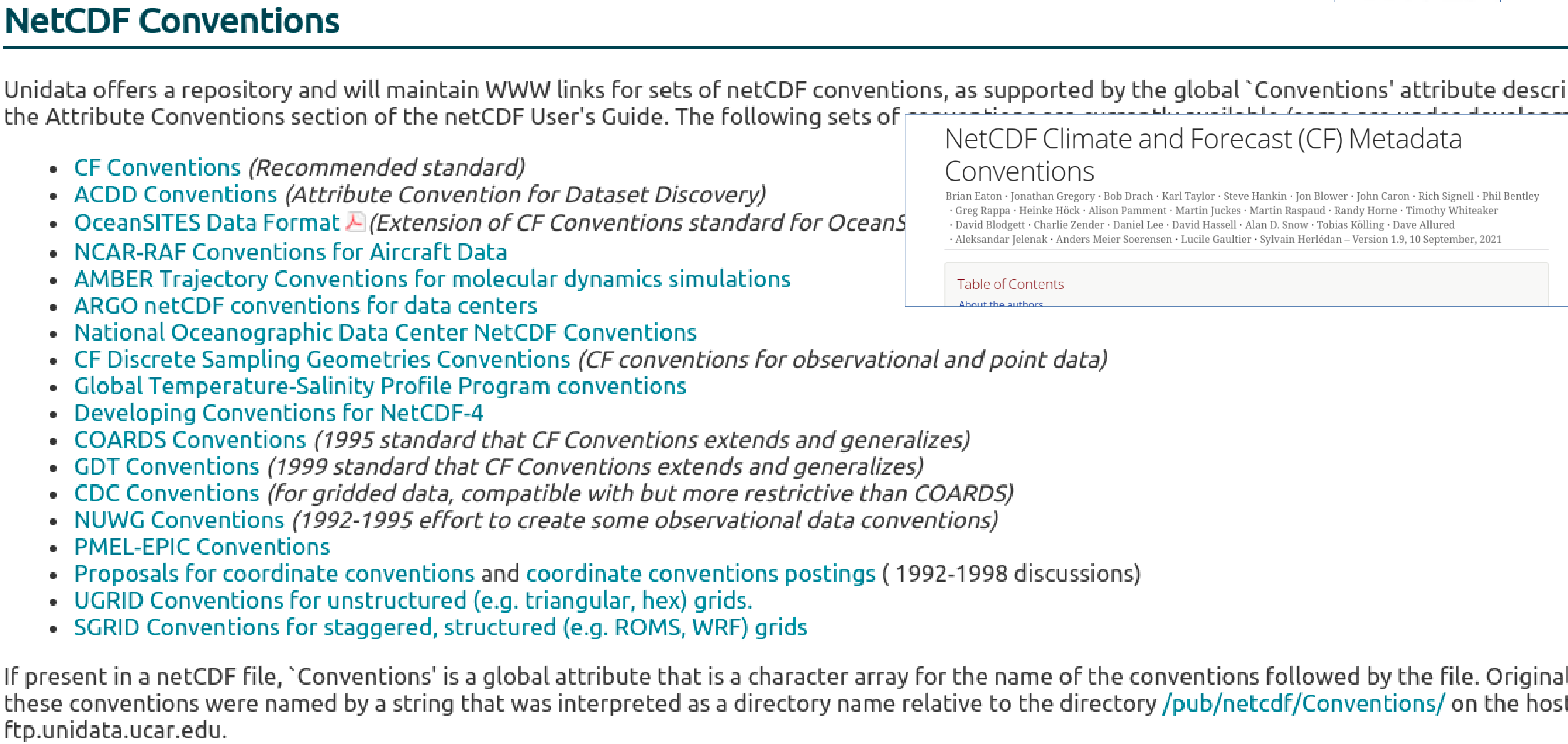
netCDF is a UCAR Unidata product, which gives CSM ready access to it's developers.

Furthermore, Unidata has responded to the requirements of CSM in terms of performance and data compression, and the resulting modifications appear in the netCDF library.

The National Science Foundation (NSF) supports the Information Infrastructure Technology and Applications (IITA) . One of IITA's primary functions is to provide funding to enhance the netC

Finally, one issue which is often overlooked but is a major concern within CSM is data management. Many different experiments will be run. The fact that netCDF is self describing means that experiment can be documented within the experiment datasets. This means that the CSM does not have to use resources (i.e., people) to maintain experiment documentation. We will be investig of data management software which supports netCDF.

NetCDF conventions



NetCDF Conventions

Unidata offers a repository and will maintain WWW links for sets of netCDF conventions, as supported by the global `Conventions' attribute descri the Attribute Conventions section of the netCDF User's Guide. The following sets of

- CF Conventions *(Recommended standard)*
- ACDD Conventions *(Attribute Convention for Dataset Discovery)*
- OceanSITES Data Format *(Extension of CF Conventions standard for OceanS*
- NCAR-RAF Conventions for Aircraft Data
- AMBER Trajectory Conventions for molecular dynamics simulations
- ARGO netCDF conventions for data centers
- National Oceanographic Data Center NetCDF Conventions
- CF Discrete Sampling Geometries Conventions *(CF conventions for observational and point data)*
- Global Temperature-Salinity Profile Program conventions
- Developing Conventions for NetCDF-4
- COARDS Conventions *(1995 standard that CF Conventions extends and generalizes)*
- GDT Conventions *(1999 standard that CF Conventions extends and generalizes)*
- CDC Conventions *(for gridded data, compatible with but more restrictive than COARDS)*
- NUWG Conventions *(1992-1995 effort to create some observational data conventions)*
- PMEL-EPIC Conventions
- Proposals for coordinate conventions and coordinate conventions postings (1992-1998 discussions)
- UGRID Conventions for unstructured (e.g. triangular, hex) grids.
- SGRID Conventions for staggered, structured (e.g. ROMS, WRF) grids

If present in a netCDF file, `Conventions' is a global attribute that is a character array for the name of the conventions followed by the file. Original these conventions were named by a string that was interpreted as a directory name relative to the directory /pub/netcdf/Conventions/ on the host ftp.unidata.ucar.edu.

NetCDF Climate and Forecast (CF) Metadata Conventions

Brian Eaton · Jonathan Gregory · Bob Drach · Karl Taylor · Steve Hankin · Jon Blower · John Caron · Rich Signell · Phil Bentley · Greg Rappa · Heinke Höck · Alison Pamment · Martin Juckes · Martin Raspaud · Randy Horne · Timothy Whiteaker · David Blodgett · Charlie Zender · Daniel Lee · David Hassell · Alan D. Snow · Tobias Kölling · Dave Allured · Aleksandar Jelenak · Anders Meier Soerensen · Lucile Gaultier · Sylvain Herlédan – Version 1.9, 10 September, 2021

Table of Contents

About the authors