# NORS as an in-situ CAMS (GAS) component: integration and compliance with CAMS in-situ component

| | |
|---|---|
| Deliverable title | NORS as an in-situ CAMS (GAS) component |
| Deliverable number | D10.4 |
| Revision | 01 |
| Status | Final |
| Planned delivery date | 30/11/2014 |
| Date of issue | 07/12/2014 |
| Nature of deliverable | Report |
| Lead partner | BIRA-IASB |
| Dissemination level | Public |

*This work has received research funding from the European Community's Seventh Framework Programme ([FP7/2007-2013]) under grant agreement n°284421.*

# DOCUMENT PROPERTIES

| | FUNCTION | NAME | DATE | SIGNATURE |
|---|---|---|---|---|
| LEAD AUTHOR | NORS coordinator | M. De Mazière | 29/10/2014 | |
| CONTRIBUTING AUTHORS | S&T partner | S. Niemeijer | | |
| | NILU | A.Vik | | |
| | BIRA-IASB research associate | A. De Rudder | | |

# Introduction

This deliverable entitled 'NORS as an in-situ GAS component' reports on the integration of NORS in the in-situ component of the Copernicus Atmospheric Monitoring Service (CAMS), and on the compliance with CAMS requirements as to data policies and metadata standards.

The exploitation of NORS results in MACC/CAMS and the integration of NORS in MACC-III VAL subproject have already been discussed in the deliverables D10.2 and D10.3, entitled 'NDACC status report' and 'NORS capacity and sustainability': NDACC data are now systematically used for the validation of MACC-II/-III products – similarly to other in-situ data – and NORS validation activities are now an integral part of the MACC-III VAL subproject and an accepted candidate for continuation in CAMS.
The way that MACC/CAMS deals with in-situ data that are archived in a central data archive, as is the case for NORS/NDACC, is that it creates a data stream between the native data archive and the MACC/CAMS processing system.
Important aspects for enbling this data stream are (1) access rights and interoperability aspects, and (2) metadata descriptions.

We will discuss briefly both aspects.

# 1. Access rights and interoperability aspects

It has been agreed from start that the NORS Rapid Delivery data are open access.
The NDACC consolidated data can be protected by the PI during maximum 2 years, but there is a general tendency to make the data publicly accessible sooner: among 197 NDACC PIs, only 18 prefer private access (J. Wild, private communication). Moreover, access can be granted to a user for specific agreed purposes, as has been the case in the past for satellite validation purposes.
Under impulse of ESA, a DCIO initiative (Data Center Inter-Operability, Meijer Y., et al, Living Planet Symposium, 2010) for exchange of catalogue information and data files between datacenters has been discussed and tested, and NDACC has participated from the start in those efforts.

# 2. Status of metadata for atmospheric observational data in MACC / CAMS

The status of metadata descriptions for in-situ data in the current MACC-II and MACC-III projects is as follows (see MACC-II Deliverable D_12.5 'Metadata for Air Quality and Atmospheric Composition" and MACC-II Delievrable D_12.6 "MACC-II recommendations for metadata description of In-situ data"):

"Although metadata implementations for in-situ observational data from scientific and regulatory data providers are ISO and INSPIRE compliant, a comparison shows that there are

large differences in the metadata implementations currently used …". In particular, some differences are justified by the different purposes of the data, e.g. the differences between regulatory and scientific data, and the fact that data that are relevant for one community are not necessarily relevant for another community.

This conclusion was drawn, among others, as a result from a Workshop on "Metadata for Air Quality and Atmospheric Composition" in Dublin from September 5 to 7, 2012, that was also attended by the NORS coordinator, who presented the GEOMS metadata as used in NORS.

Hereinafter, some comparisons between the NORS GEOMS and other current metadata standards are briefly commented:

## 2.1. GEOMS versus INSPIRE and other metadata standards.

### 2.1.1. GEOMS versus INSPIRE

During that workshop, it became clear also that there are a number of obstacles in mapping scientific metadata for atmospheric composition to the INSPIRE metadata regulation; the one that is most applicable to the domain of atmospheric composition is the one for the data theme 'Atmospheric Conditions and Meteorological Geographical Features ' but it still requires some revision before being applicable. Among the problems is the fact that the INSPIRE regulations are strongly focused on administrative data like cadaster, elevation, etc. The full implementation is aimed at the year 2020.

It is currently not 100% clear if there are any conflicts between the GEOMS and INSPIRE initiatives, but it is not likely. It is actually unlikely that INSPIRE will have any influence on the GEOMS standard or its data centres at all. The GEOMS data are research oriented data and it is mainly measurements of non-surface atmospheric properties. INSPIRE, on the other hand, is about cartography and it aims to establish electronic map-services (spatial infrastructure) which are compatible across borders in Europe.
INSPIRE is all about services and interfaces, and it does not regulate how a data set is formatted or what internal standard a data centre has.
Having that said, it is still possible that some of the data being handled by the GEOMS data centres will have to relate to INSPIRE one way or the other.

*Conclusion*
It is useful to keep an eye on the developments and implementations of the INSPIRE directive, but it is not likely that it will affect the type of data that the GEOMS standard is dealing with.

### 2.1.2. GEOMS versus netCDF-CF

Another standard for metadata descriptions that is often used in the atmopsheic modelling community and in the satellite community is the netCDF-CF standard.

We have made an effort during the NORS project to make a mapping between the GEOMS and the netCDF-CF metadata.

A preliminary incomplete mapping is provided in Annex to this deliverable.

## 3. How does NORS/NDACC comply with the recommendations formulated in MACC-II as to metadata descriptions

The following recommendations as to metadata descriptions were formulated in MACC-II:

- Develop mapping tools between existing standards

  ⇨ This work has been initiated in NORS as outlined in Section 2.1.

- Develop a common controlled vocabulary, e.g., for data quality flags, for variable names, for identification of compounds

  ⇨ An effort towards a common vocabulary with netCDF-CF was made when defining GEOMS variables names. If needed, this effort will be pursued and – when names differ – a mapping tool for variable names could be developed.

- Secure continuity and impact of metadata harmonisation activities; more specifically, Proceed with joint working groups, link to WMO Working Groups under the GEO Work Plan and to INSPIRE

  ⇨ Such links are secured by NORS/NDACC partners and will be pursued.

# Annex

# 1 Mapping of GEOMS vs netCDF-CF

This document provides a first assessment on a mapping between the GEOMS and netCDF-CF metadata standards.

| *Topic* | *GEOMS* | *netCDF-CF* | *Comments* |
|---|---|---|---|
| File format | HDF4, HDF5, and netCDF<br><br>Data types restricted to: float32, float64, int16, int32 and char (i.e. string).<br>Floating point values should follow IEEE 754.<br><br>There are restrictions on what to use in HDF4, HDF5 and netCDF. | netCDF only<br><br>Data types restricted to those of netCDF. | netCDF-CF can theoretically also be applied to HDF4 and HDF5. The metadata is in both cases portable enough to be brought to another data format.<br>Also note that netCDF4 is actually HDF5.<br><br>Note that GEOMS does not support an (u)int8 data type. |
| Global Attributes | Tries to describe full information as needed for a metadata catalogue. Value space is generally very strict. Consists of originator, dataset, and file attributes. | Only provides a few basic attributes. Value space is generally undefined. | See Section 1.1 |
| Variable Attributes | | | See Section 1.2 |
| Naming convention | Names are composed of a 'name' and optional 'mode' and 'descriptor' parts. The 'mode' distinguishes the 'how' between variables with the same name. The 'descriptor' distinguishes related parameters from primary parameters (e.g. for error information) | | |
| Specification of ranges | | | When to use low, high, middle, and/or length? |
| Time format | MJD2K for variables values and ISO 6801long format for global attribute values.<br><br>Leap seconds in variable values handled using epoch shift equal to introduced leap | | |

| | | | |
|---|---|---|---|
| | second.<br><br>No specific calendar support | | |
| Axis | VAR_DEPENDS references itself.<br>Axis variables can be two dimensional (if they change over time). | | |
| Coupling of dimensions | VAR_DEPENDS refers to VAR_NAME of axis variables for dimensions (or INDEPENDENT).<br>Scalar variables use CONSTANT. | | |
| Units | udunits with a few extensions. Most notably 'MJD2K' to define the unit for time and 'molec' in e.g. molec/cm^3 for concatenations. | udunits and strictly sticking to SI units. | For example, netCDF-CF will use mole/cm^3 instead of molec/cm^3 for concentrations. |
| Geographical coordinate system | Limited to lat/lon grids (e.g. WGS84) | | |
| Characterset | ISO 646/US ASCII | | |
| Dimension ordering | time, latitude, longitude, altitude, wavelength, ...<br>Ordering of dimensions in attributes uses C convention. | time, altitude, latitude, longitude, ... | C convention vs. Fortran convention. |
| Templates | Templates exist that define strict subsets of the standard (similar to 'product types' in the EO domain). Templates are defined using a custom CSV format. | Templates are not part of the standard.<br><br>Format definitions are sometimes specified using the CDL format. | |
| Compression by gathering | Not supported | A sparse multi-dimensional array can be compressed. | If we would have X(x), Y(y), and Z(x,y) this can be stored as X(k), Y(k), Z(k), K(k). NetCDF-CF then adds a 'coordinates' attribute to Z with value "X Y" to indicate that Z is actually meant to be a 2D array and a 'compress' attribute to K with values "x y" (i.e. the names of the netCDF dimensions containing the original lengths of the x and y axis). K contains a flat index in the x*y array for each point. |

## 1.1 Global Attributes

| GEOMS | netCDF-CF | Comments |
|---|---|---|
| PI_NAME | | |
| PI_AFFILIATION | | |
| PI_ADDRESS | | |
| PI_EMAIL | | |
| DO_NAME | | |
| DO_AFFILIATION | institution | |
| DO_ADDRESS | | |
| DO_EMAIL | | |
| DS_NAME | | |
| DS_AFFILIATION | | |
| DS_ADDRESS | | |
| DS_EMAIL | | |
| DATA_DESCRIPTION | title (netCDF generic) | |
| DATA_DISCIPLINE | | |
| DATA_GROUP | source | |
| DATA_LOCATION | | |
| DATA_SOURCE | | |
| DATA_VARIABLES | (implicit) | |
| DATA_START_DATE | | |
| DATA_STOP_DATE | | |
| DATA_FILE_VERSION | | |
| DATA_MODIFICATIONS | | |
| DATA_CAVEATS | | |
| DATA_RULES_OF_USE | | |
| DATA_ACKNOWLEDGEMENT | | |
| DATA_QUALITY | | |
| DATA_TEMPLATE | | |
| DATA_PROCESSOR | | |
| FILE_NAME | | |
| FILE_GENERATION_DATE | | |
| FILE_ACCESS | | |
| FILE_PROJECT_ID | | |
| FILE_ASSOCIATION | | |
| FILE_META_VERSION | Conventions (netCDF generic) | |
| FILE_DOI | | Digital Object Identifier |

| | | |
|---|---|---|
| | source | |
| | history (netCDF generic) | |
| | references | |
| | comment | |

## *1.2 Variable Attributes*

The HDF4, HDF5 and netCDF formats also intrinsically contain information regarding the stored variables. What is described in GEOMS sometimes overlaps with this internal information. The table below shows both the relation of GEOMS variable attributes towards those of netCDF-CF as well as to the metadata intrinsically maintained by a data format.

| GEOMS | HDF4 | HDF5 | netCDF | netCDF-CF |
|---|---|---|---|---|
| VAR_NAME | (implicit) | (implicit) | (implicit) | standard_name |
| VAR_DESCRIPTION | long_name | | long_name | |
| VAR_NOTES | | | | comment |
| VAR_SIZE | (implicit) | (implicit) | (implicit) | |
| VAR_DEPEND | (implicit; optional)[1] | [2] | (implicit)[3] | axis (for axis variables) |
| VAR_DATA_TYPE | (implicit) | (implicit) | (implicit) | |
| VAR_UNITS | units | | units | |
| VAR_SI_CONVERSION | | | (automatic via udunits software) | (netCDF-CF already uses SI) |
| VAR_VALID_MIN | valid_range | | valid_range, valid_min | |
| VAR_VALID_MAX | valid_range | | valid_range, valid_max | |
| VAR_FILL_VALUE | __FillValue | (property) | _FillValue, missing_value (deprecated) | |
| | format | | | |
| | cordsys | | | |
| | add_offset | | add_offset | |
| | add_offset_err | | | |
| | scale_factor | | scale_factor | |
| | scale_factor_err | | | |
| | calibrated_nt | | | |
| | | | signedness (deprecated) | |
| | | | C_format | |

1 HDF4 has native support for named dimensions (SDsetdimname). The use of named dimensions is however optional.
2 HDF5 has no build-in support for named dimensions (or otherwise coupling dimensions of datasets).
3 netCDF has build-in support for named dimensions (nc_def_dim). Using named dimensions is mandatory for netCDF.

| | | | | |
|---|---|---|---|---|
| | | | FORTRAN_format | |
| | | | | ancillary_variables |
| | | | | bounds |
| | | | | calendar |
| | | | | cell_measures |
| | | | | cell_methods |
| | | | | climatology |
| | | | | compress |
| | | | | coordinates |
| | | | | flag_masks |
| | | | | flag_meanings |
| | | | | flag_values |
| | | | | formula_terms |
| | | | | grid_mapping |
| | | | | institution |
| | | | | leap_month |
| | | | | leap_year |
| | | | | month_lengths |
| | | | | positive |
| | | | | references |
| | | | | source |
| | | | | standard_error_multiplier |

## *1.3 Variable names*

Mapping of specific categories of variable names to be described here...