A generic database for earth observation data. ESA Campaign Data Service

A.F. Vik, T. Krognes, S. Bjørndalsæter, C. Stoll, S.-E. Walker, B. Gloslie, R. Paltiel, T. Bårde, R.L. Våler

Norwegian Institute for Air Research, NILU, P.O. Box 100, 2027 Kjeller, Norway –
(afv, tk, sbj, cst, sew, bgl, rpa, trb, rlv) @nilu.no

Abstract – The ESA (European Space Agency) Campaign Database (CDB) was developed and implemented at the Norwegian Institute for Air Research to provide the EOP-S campaign section of ESRIN-ESA with sufficient support for storage of campaign data. The CDB builds on the experience of the ENVISAT Cal/Val database, but has been further developed to handle data from all ESA campaigns and from a multitude of earth observation sciences. The database, and its underlying structure, is suitable for archiving and indexing all forms of geophysical/geo-located data in a common context. CDB is available through http://nadir.nilu.no/cdb.

Keywords: Database, Campaigns, Cal/Val, ENVISAT, EO.

1. INTRODUCTION

The ESA (European Space Agency) Cal/Val database was developed and implemented at the Norwegian Institute for Air Research to provide ENVISAT scientist with a common framework and repository for exchange of correlative data, mainly from ground based measurements. The experience from this activity led to a new ESA initiative to develop a more generic database, the ESA Campaign Database (CDB). This system is a generalisation and further development of the Cal/Val system used for some ENVISAT calibration and validation campaigns. Differences are kept to a minimum, to make the transition easy for the user community of the original system. The CDB includes all data and metadata definitions from the previous Cal/Val data centre, but is able to handle data from all ESA campaigns. It is a system for storing and indexing complex data sets from a multitude of earth observation sciences, and is no longer a database for correlative data only. The database, and its underlying structure, is suitable for archiving and indexing all forms of geophysical/geo-located data in a common context.

The objective of the campaign database is to provide an online information system that supports users in managing and exploiting campaign datasets for Earth Observation missions and applications. In a more future perspective the overall aim is to provide a data centre that handles Cal/Val data, satellite data and campaign data in an integrated way. The centre will in this way increase the dissemination potential for all classes of data. The database is built with a strict quality control of incoming data and options for individual file-formatting is very limited. Using the same principles also for non Cal/Val data, will simplify the use of multi disciplinary data since all files are part of the same uniform data set.

The ESA Campaign Database is available at http://nadir.nilu.no/cdb, and a user account is required to enter the restricted part of the data centre. The account is personal and will give a user access to data from one or more campaigns.

2. DESCRIPTION OF THE CAMPAIGN DATA SERVICE

2.1 Description of metadata

Metadata are in fact data about data. They provide the information that the data-user needs in order to understand the actual data. For an atmospheric observation, the data can be a series of numbers that does not make sense unless you provide the metadata on what the numbers represent. Typical metadata in this case would be time and location of the measurement, what parameter is measured, what is the uncertainty in the measurement, who did the measurement, what unit is used, etc.

In the ESA project on calibration and validation of ENVISAT (Cal/Val), a comprehensive effort was put down into developing a structure for defining such metadata (Bojkov et al., 2002). The structure was based on previous developments at NILU, mainly through the experiences gained from the EMEP database that has been operative since 1979. The structure of the Cal/Val database specifies all the metadata parameters that are needed for each data file. Table 1 shows the complete list of metadata parameters used in HDF files at the ENVISAT Cal/Val data centre. The structure is very flexible and is designed to store most types of measurements. The first entry in the table is in fact 12 different parameters that are used to identify the owners of the file. The Variable description and Visualisation attributes must be separately declared for each variable, while the other attributes only occur once in the file.

In addition to this structure, most metadata parameters are associated with a separate list of legal values that must be used to describe the observation. This makes it easier to store similar or related types of observations in a comparable manner. As an example, a variable containing ozone measurements should be named O3.CONCENTRATION, and not ozone, ozone_concentration, etc. Only the legal values of metadata will be accepted by the database.

CDB builds on the efforts laid down in the ENVISAT Cal/Val project and reuse the same lists of legal parameters. However, new entries to these lists have been provided in order to cope with the different requirements and scope of CDB. This work is furthermore a continuous effort since new campaigns are regularly starting to use the data centre. Because of this, a lists of continuously updated legal values of the various metadata parameters are provided on the database web-interface. In addition, a complete description of the rules behind all metadata definitions and lists of all legal values are available to the users in a metadata guidelines document.

Table 1. Metadata parameters used in the CDB database.

Originator Attributes
PI, DO, DS with NAME, AFFILATION, ADDRESS and EMAIL
Dataset Attributes
DATA_DESCRIPTION
DATA_DISCIPLINE
DATA_GROUP
DATA_LOCATION
DATA_SOURCE
DATA_TYPE
DATA_VARIABLES
DATA_START_DATE
DATA_FILE_VERSION
DATA_MODIFICATIONS
DATA_CAVEATS
DATA_RULES_OF_USE
DATA_ACKNOWLEDGEMENT
File Attributes
FILE_NAME
FILE_GENERATION_DATE
FILE_ACCESS
FILE_PROJECT_ID
FILE_ASSOCIATION
FILE_META_VERSION
Variable Description Attributes
VAR_NAME
VAR_DESCRIPTION
VAR_NOTES
VAR_DIMENSION
VAR_SIZE
VAR_DEPEND
VAR_DATA_TYPE
VAR_UNITS
VAR_SI_CONVERSION
VAR_VALID_MIN
VAR_VALID_MAX
VAR_MONOTONE
VAR_AVG_TYPE
VAR_FILL_VALUE
Variable Visualisation Attributes
VIS_LABEL
VIS_FORMAT
VIS_PLOT_TYPE
VIS_SCALE_TYPE
VIS_SCALE_MIN
VIS_SCALE_MAX

2.2 Description of database architecture and functionality

NILU has designed and implemented a system for organizing ground based measurement data, and for retrieval of the same data by scientists that perform comparisons with measurements from the ENVISAT satellite. The work has been performed in close co-operation with ESA and with representatives of the user community. The system is complex since it entails co-operation between wide spread scientific communities that have separate and different cultures and methods. In the ESA ENVISAT Calibration/Validation effort the measurements of stratosphere physicists, modellers and mathematicians, marine biologists, and space scientists needed to be described within one common frame of reference. As the system evolved through a generalisation process into CDB, an even larger user community had to be incorporated. The system that handles this task is described in the following.

The HDF 4.1r3 file format was chosen for data exchange, based on the established use of this format within ESA and some of the user groups. The system furthermore handles archiving and sharing of documentation, images and data in any file-format for special purposes. Main software tools have been developed in FORTRAN, IDL, Macromedia CF (Cold Fusion) and Flash, and UNIX shell-scripts. The system uses Red Hat Linux, Apache web-server with CF server-side scripting, and a MySQL database.

The system components are here described in a logical order when we follow a data file as it passes from the originator into the storage and forward to an end user. Fig. 1 shows a schematic diagram of how the various components are connected.

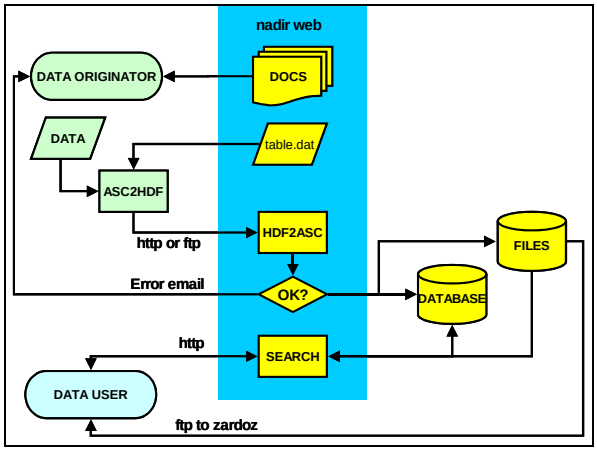

Figure 1: Schematic diagram of data flow from data originator (green modules) to file collection and index database and back to data user (blue module).

The DS (Data Submitter) needs to sign a data protocol and be registered in the system metadata. This allows the user access to the CDB web site, and gives permission to upload data for one or more campaigns and sub-projects (AO's).

At the CDB website the user will find a database manual, file templates and other documents that help with formatting original data into an HDF file. A software tool named ASC2HDF is available for Windows, Linux, Solaris and HPUX users. This tool accepts data and metadata in two simple text files, and will generate an HDF file after extensively testing the input. A user-friendly front-end to the ASC2HDF program has recently been developed, and this allows the user to fill in data in an excel sheet and to save the

content directly as a HDF file. The tool ensures that the user formats the data correctly.

When the HDF file has been successfully tested at the local site, it may be uploaded to the CDB site by ftp, or through a web upload page. File processing scripts (HDF2ASC) check for new files every 5 minutes. Even files that have been successfully tested by the originator, may be rejected at NILU, mostly due to inconsistencies in the file name (which reflects a subset of the metadata content), or due to duplicate file names or out-of-sequence version numbers. If the data supplier is not accredited for the campaign or sub-project listed in the file, the file will also be rejected. An error report will automatically be emailed to the data supplier and the owner of the logon name that was used, and the file will be moved to a hidden directory. If all checks out correctly, the received HDF file will be moved to a storage file tree, and the file name, upload details and central metadata elements are stored in an index database. The system enforces consistent naming of variables and other metadata elements, and consistent spelling of names for people, organisations and sites.

The index database contains the official list of allowed metadata values in the CDB HDF data files, in addition to logs of uploaded/downloaded files, an overview of metadata contents, and the variable list of all accepted HDF data files. All this information is available to dynamic web pages at the web site. The main end user tool on this site is the "Search Data" page, which allows the user to sort through the data files with advanced criteria selections. Filtering by data supplier, project, location, data source, data type, component and other metadata elements is supported. Data files may also be filtered by a "4-D box algorithm" (any file with data relevant for a given geographical location and time). Furthermore, files can be filtered by submission date and update status. All data files that match the search criteria are listed in a new web page, with links to HDF data file download, to comments, and to a variable list. In the variable list page the user may select variables and generate an on-line plot. Plotting is supported for up to 4D data arrays. An example of plotting of a 2D data array is shown in Fig. 2.

In the file list the user may also select multiple files for download as a tar-ball. The user may save the search criteria in the index database for convenient re-use at a later time.

In addition to the described search interface a graphic interface showing the geo-graphical location of the data may be used for searching. This is a webpage with a Macromedia Flash MX application showing a world map with information on geo-location. The map works in two modes, the Station and the Trajectory mode, and the user may zoom into the map to study details. The station mode displays data where geo-location is constant within a data file and the trajectory mode displays data where geo-location varies over time, but is constant for each time step in the file. The trajectory mode furthermore displays the altitude of the location by a colour scheme. A user may retrieve data by clicking on a dot (in the station mode) or on a trajectory. Within the flash-application a filter-tool with drop-down menus similar to those of the already described text-based search interface is available, and the user may chose one or several parameters. The map in the background will automatically be updated when the user chooses a value for one of the parameters and dots or trajectories not matching the chosen values will vanish. This allows the user to differentiate between all available data files and the user will see the location of all data files before he/she press the "Get file(s)" button at the bottom of the filter. It is also possible to click on any of the remaining dots/trajectories to retrieve data from only that parcel or location. A screen shot of the mapping tool is shown in Fig. 3.

Users that have an IDL license may download IDL scripts for HDF data file formatting (excluding the detailed error checking available in the FORTRAN version) and for plotting of data sets from HDF files.

A new feature of CDB was the implementation of Project Internal Pages (PIP), where users may share campaign specific information through a web-portal. The PIP contains sections for documents, a link archive, contact information, an image gallery and a discussion board. A user account is also needed to access these pages.

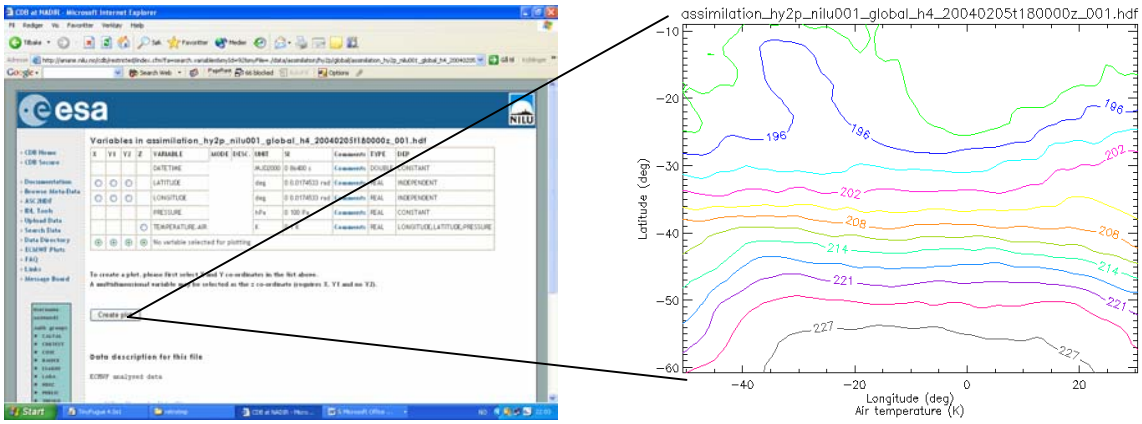

Figure 2: A screenshot from the CDB web pages. Metadata on variables can be browsed on-line, and data may be visualised through built-in plotting routines.

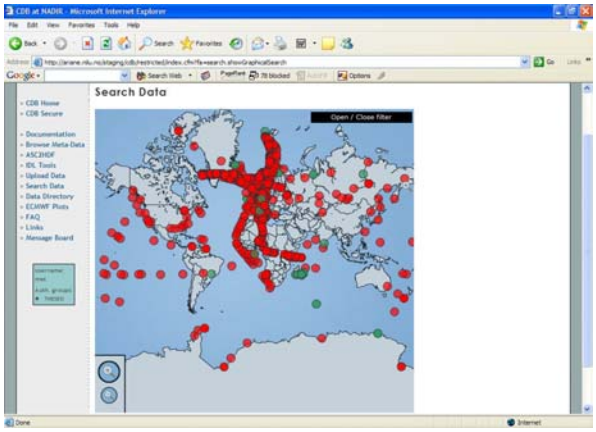

Figure 3: Screenshot of the mapping tool in the station mode. The user sees the location of all files in the database, but will only have access to those marked with green.

2.3 User support and archiving strategies

CDB aims to increase the use of geophysical data after a campaign is completed. Measurements are made available for other scientists (only after permission is given from original PI) and data are no longer sent to rest in the drawer of a scientist desk. CDB provides the final archive for the data. Another advantage with using the CDB is the possibility of sharing data within the campaign consortium – both during the campaign and in the analysing phase. An important part of the CDB operations is therefore provision of user support and advisory to data managers on how to archive data and what types of data they should store.

All registered users have access to on-line information through the web portal and also to a CDB handbook that explains how to use data and how to format, upload and manage files. A help-desk is available to all users of the data service and technical personnel are available to solve file formatting and system related issues to data submitters and data users. Furthermore, there are scientific personnel available to solve less technical issues and to assist data users with interpretation of data and to provide campaign data managers with assistance on archiving strategies.

A campaign data manager is commonly involved in a campaign to work closely with the scientific campaign coordinator and will thereby have the general overview of all the data collected through the campaign or project. For CDB, the main task of the data manager is to set down guidelines for reporting of data so that individual data submitters know what data that should be archived and how to format their files. The campaign data managers furthermore needs to look beyond the scope of their campaign and try to see how archiving of the campaign data fit in with the objectives of CDB.

The metadata guidelines (as described briefly in section 2.1) define rules for what names and values that can be inserted in a data file. Apart from this, it provides no rules for how the data structure should be defined in a file. As an example, for ozone sondes it is possible to store the ozone concentration (mPa) as a function of altitude, total pressure, time after launch, etc. A template provides the data submitter, i.e. the person responsible for creation and upload of data files, with a guideline for how he or she should define independent and dependent variables and which of these it is necessary to include. An archiving strategy is implemented in order to keep the different templates compatible, so that there is a uniform way of archiving data from different measurements and platforms. This allows for easier comparisons of different observations.

Before a campaign is performed, the PIs or campaign manager have clear goals for what they want to achieve with the measurements. For the ESA campaigns to be stored at CDB, it is necessary to keep these goals in mind when data are to be converted into HDF files and stored at the data centre. It is furthermore important to keep in mind that CDB is a database for several campaigns, and data from one campaign could be used by other campaigns (please note that data are not automatically shared between campaigns, and that sharing of data only occurs after an agreement with ESA and campaign managers). Such reuse of data may justify upload of more or other types of data.

There is also a question regarding the level of data to be archived. The only requirement regarding presence of variables in the CDB data files is that they must contain a time and geo-reference. Data must therefore be of level 1 or higher. A measurement campaign normally includes further processing of data into physical values such as ice thickness, gas concentration, vegetation index, etc. When reporting data to CDB, the data manager must again consider the purpose of the campaign and choose what levels of data that should be archived. Sometimes, it is beneficial to store all campaign data products, on all available levels. This will allow for future reprocessing of data sets.

Documentation of several example campaigns and methods on how to develop a sustainable archiving strategy is available for all users of CDB.

3. ACKNOWLEDGEMENTS

The authors thank the European Space Agency, ESRIN and ESTEC divisions, for support to the developments and operations of the Campaign Data Centre.

4. REFERENCES

Bojkov, B.R., Mazière, M. and Koopman, R.M., “Generic Metadata guidelines on atmospheric and oceanographic datasets for the Envisat Calibration and Validation Project”, European Space Agency, ESRIN, Frascati, Italy (2002).