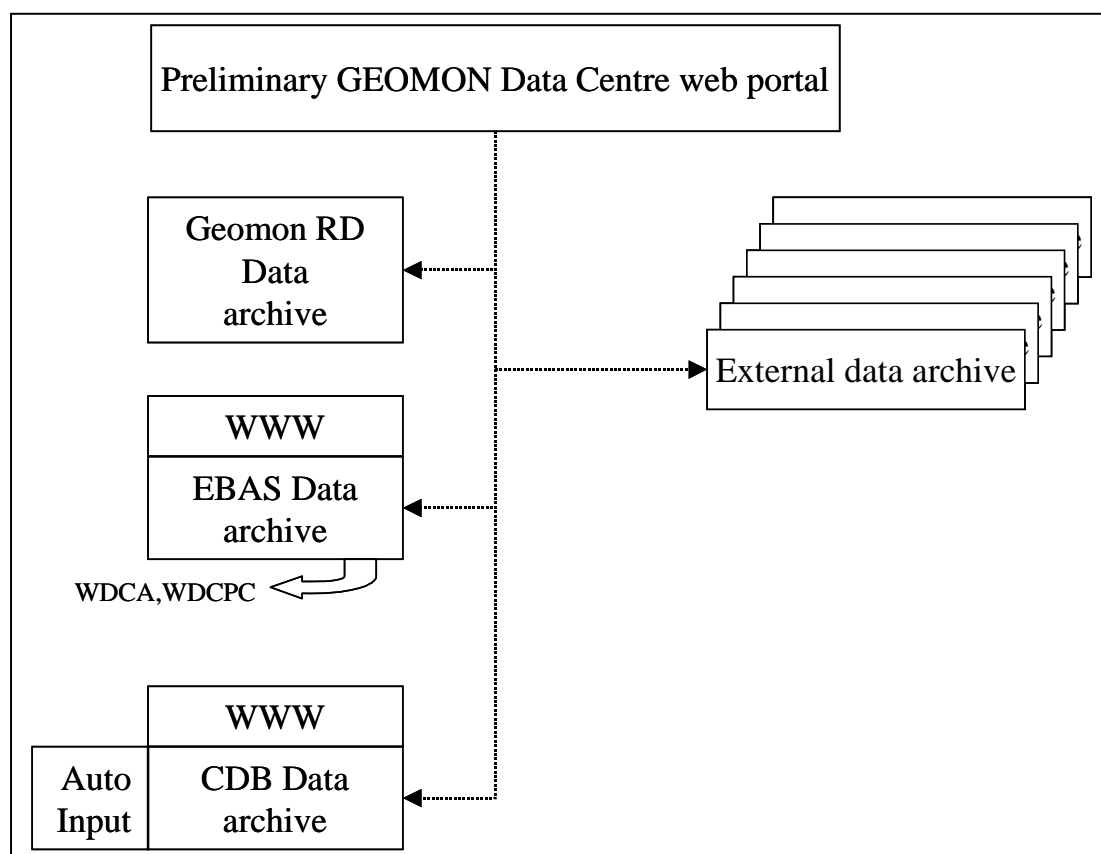


Planned developments of the GEOMON data centre

Based on previous discussions, the draft GEOMON data centre concept paper, and the data flow diagrams, we now suggest to establish the data centre by building on existing infrastructures with a long-term perspective on operations. As the data flow of the project is not yet completely defined, some of the proposed suggestions below may need further refinement.

Current state of the GEOMON data centre

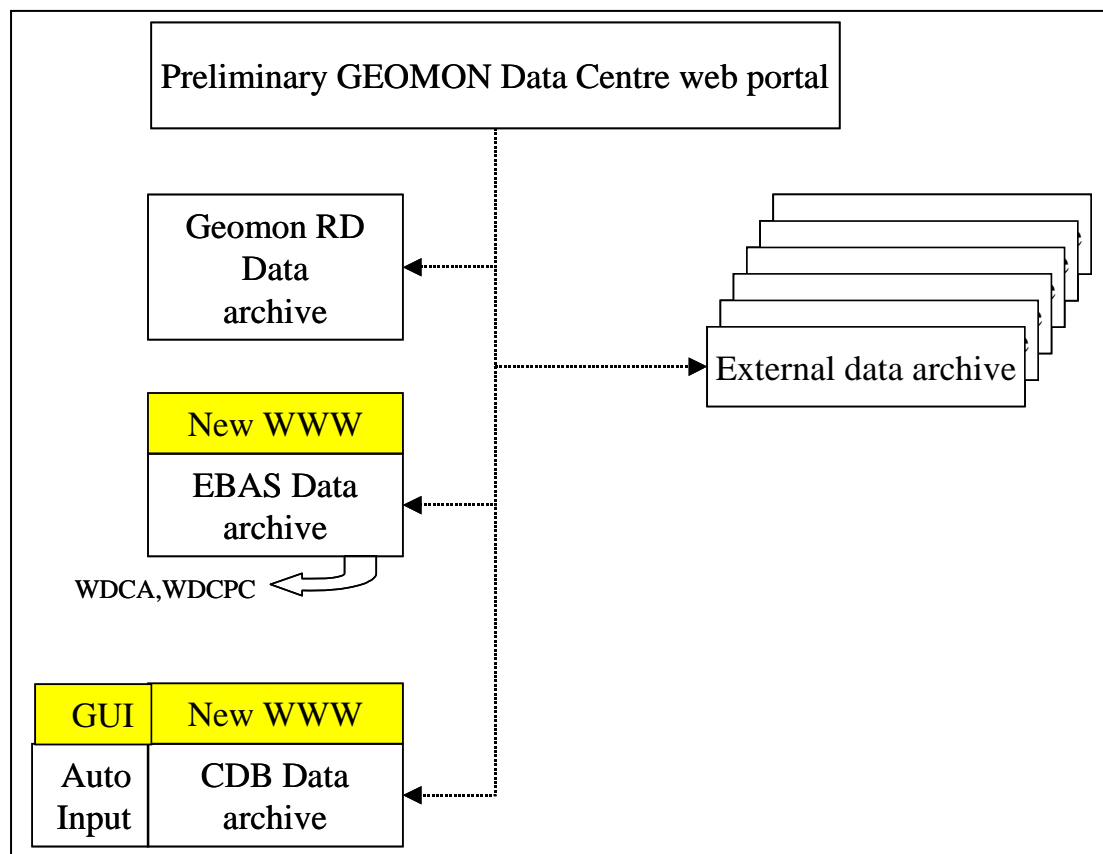
The current state of the data centre is presented in the figure below:



A preliminary web portal has been set up at NILU. This contains static links to external data archives that contain data that are either supposed to contain GEOMON funded measurements or contain data that are of direct interest to the objectives of the project. In addition, the preliminary web portal provides static links to three data archives hosted at NILU: a) the project Rapid Delivery Data archive that host preliminary data, b) the EBAS data archive, with its current web portal (that contains static links to data content) and c) the ESA Campaign Database (CDB) with its current web portal that allows online search and retrieval of data. CDB features an automatic file processing system that checks all incoming data for consistency and rejects files that are incorrectly formatted. Official data sharing agreements exists between EBAS and WDCA (GAW, World Data Centre for Aerosols) and WDCPC (GAW, World Data Centre for Precipitation Chemistry). EBAS will be used for groundbased in-situ measurements in general, while CDB is intended for aircraft measurements and remote sensed data.

Ongoing developments

The following improvements of the GEOMON data centre and its related components are currently ongoing.



A graphical users interface to make it easier to work with HDF files

The ESA CDB standard file format is HDF (using the AURA and Envisat Cal/Val metadata standard) and NILU provides users with a standard tool ASC2HDF to help users correctly format their data files. ASC2HDF requires two ASCII input files and a simple Excel tool exists to make it easier to format these two text files. NILU is currently developing a new GUI (Graphical User Interface) to work on top of ASC2HDF that should further reduce the threshold of getting started with this file format. The GUI is expected to be completed in November 2007 and will be available for GEOMON users. The GUI will work on Windows/Linux/Solaris and will enable users to read/write/edit CDB HDF files.

Update of the CDB web portal

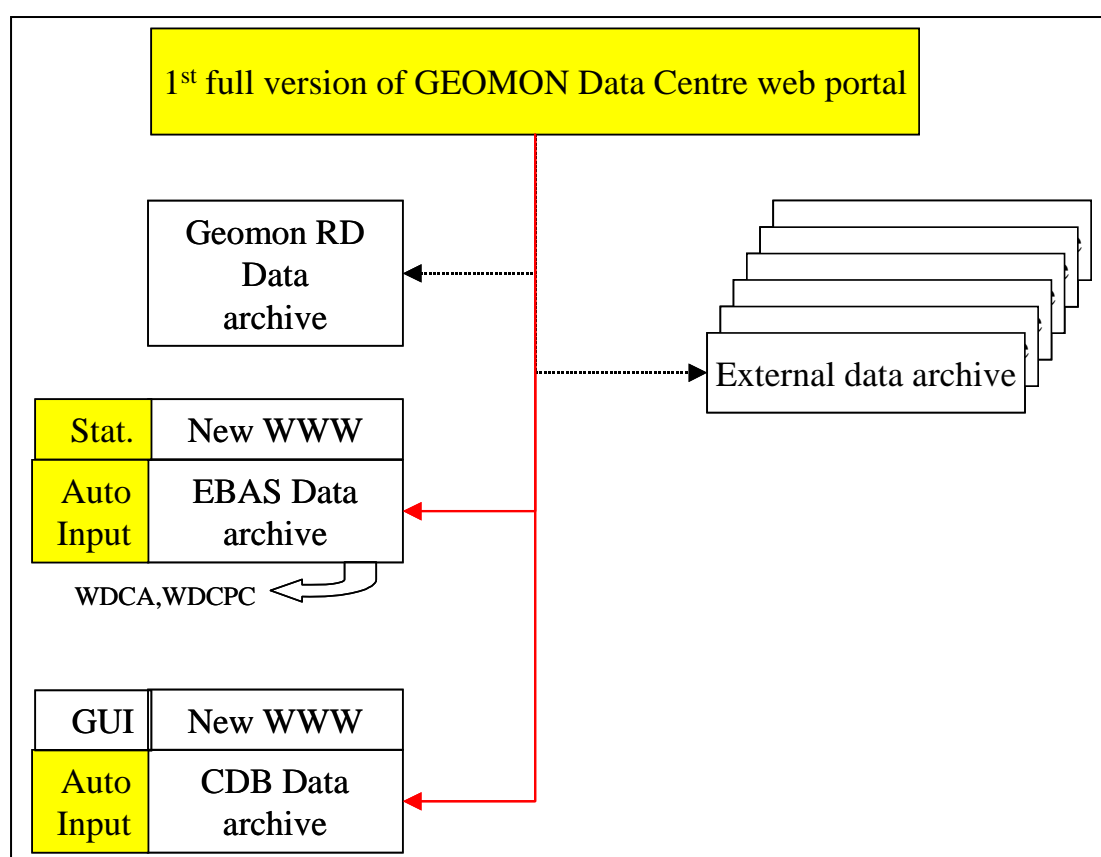
The CDB web portal is currently being updated and will allow for a better and more logic access to data. The portal is trying to provide necessary information to the different types of users (data users, data submitters and campaign managers) and tries to avoid overloading them with information they don't need. The new portal is expected to be completed in November 2007 and will contain a link to the new GUI. The previous CDB web portal is available at <http://nadir.nilu.no/cdb> (user account required).

Update of the EBAS web portal

The development of a new web portal for the EBAS data archive has been going on for some time, but NILU has not yet finalised these pages (mainly due to the development of the GUI for CDB). When completed, they will provide users with a graphical interface to search for and retrieve data from the EBAS data archive.

Planned developments

Below are suggestions for how the data centre could evolve from its current state. Specifications on how the data centre is supposed to further develop will be revised in month 18 of the project, and the current plan should be seen as a draft. Comments and suggestions for modifications are very welcome.



Stat. refers to "Statistical Tools", described below.

Automatic File ingestion in EBAS

The current procedure of ingesting data to EBAS consists of users sending data files to NILU on email or uploading data to ftp.nilu.no. NILU scientists then manually check file format and insert the data into the database. Through the EUSAAR project, an automatic file processing unit, similar to that in operation for the CDB system, will be developed. This will simplify the data upload procedure significantly. The developments of the file processing unit will start in November 2007 and should be available around March 2008.

1st full version of GEOMON data portal

The current web portal to access GEOMON data and related data sets should be seen as preliminary and will as soon as possible be replaced by a new web portal that enables search for and retrieval of GEOMON data (in this context, GEOMON data are those funded by or being relevant to the project) through a single search interface. In its first version, the data portal will allow users to search for data in EBAS and CDB at the same time, but links to RD data and externally archived data will remain static (this means that it will not yet be possible to find data stored in external archives through a common unified search interface. User will instead have to follow static links to external archives and search for data there). The developments of the search interface will commence early 2008 and should be available around March 2008. The system will be designed to allow further expansion in the future and it is a defined goal to enable search for any GEOMON related data (not just those stored at NILU) through this interface. This will, however, require implementation of machine-to-machine interfaces (to be accomplished later in the project).

The GEOMON data portal will have the GEOMON logo on top of the page, but will contain logos of all participating networks and projects. Providing visibility to all participating scientist and programmes is seen as essential for the success of such a portal. Data origin and ownership must be part of metadata.

Definition of metadata/data content and format for GEOMON data that goes into EBAS/CDB

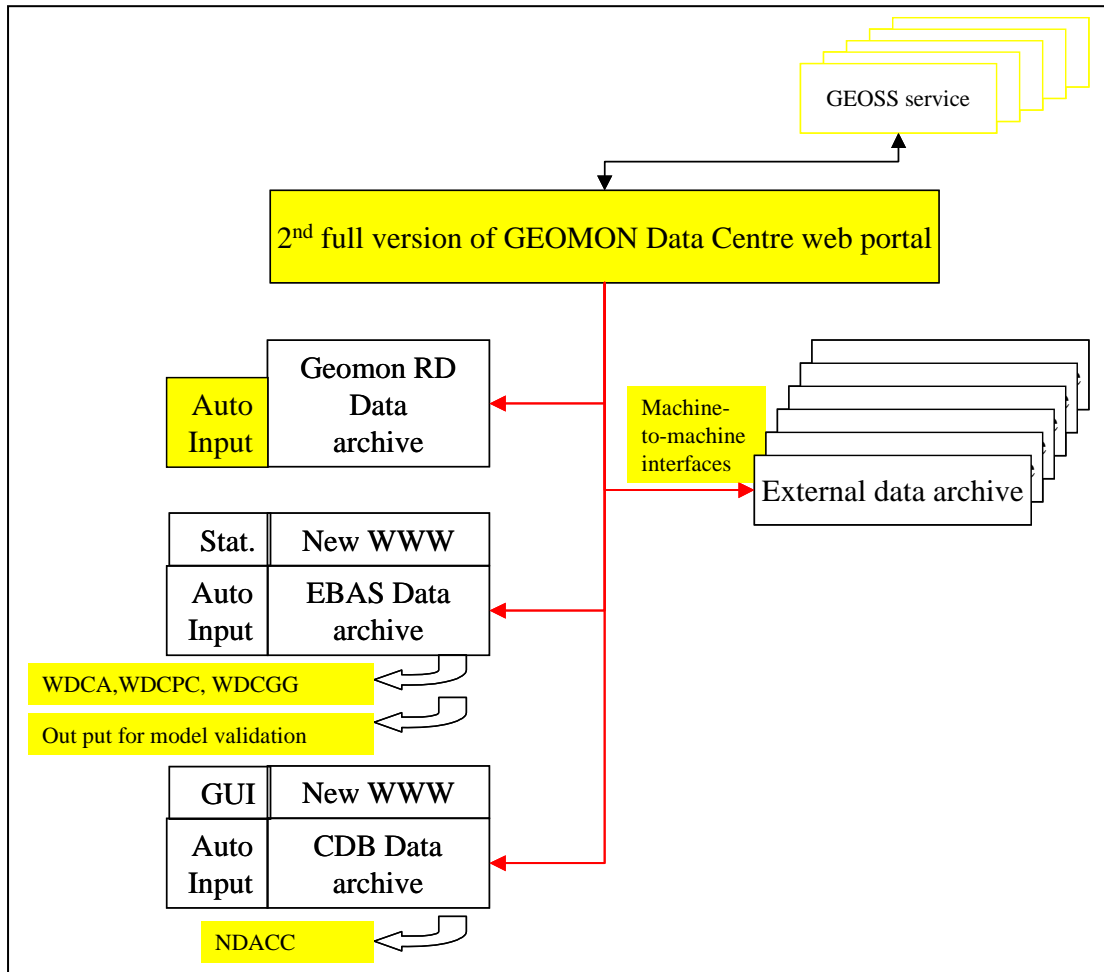
This is not directly a system upgrade, but it concerns the specifications for data formats and data file content that should be defined for all measurements that are to be uploaded to the data centre. EBAS and CDB have limited flexibility regarding their input data formats, and it should be discussed if other data formats should be supported. This will then possibly require developments on the two data systems and especially on their file processing units.

Implementation of statistics tools for EBAS

While CDB is in fact a meta-database providing a searchable index of a large file-archive, the EBAS system is a database where all the numbers are actually stored in the indexed applications. It is therefore difficult to perform large statistical operations with data in CDB (since data files would have to be read and interpreted first), but the EBAS system is suitable for statistical applications calculating averages, percentiles, standard deviations, etc. Such tools are already available at NILU, but they will also be made available through the EBAS and GEOMON web portals in 2008. This is done as part of the EUSAAR project, but will also benefit the GEOMON project.

Further plans for developments (less specific on timing) in GEOMON

In addition to the plans listed above, are plans for developments that will come later on in the project. These are to be specified in more detail around month 18 of the project. All the below improvements are to be provided as part of the 2nd full version of the GEOMON data centre web portal



Export interface for model validation

A special interface will be developed for EBAS (and, if necessary also on CDB) to allow modellers exporting data out in formats suitable for automated model validation. The nature of this interface is to be specified through the HTAP (Hemispheric Transport of Air Pollution) activity where NILU is leading the task on providing access to observations. The interface will most likely support the CF-conventions and provide data in the netCDF format.

Machine-to-machine interfaces to external data archives that are not hosted by NILU

In order to support searching in data archives that are not hosted by NILU, it is necessary to establish links to the external systems so that the search interface has an overview of all the GEOMON data (but these data are possibly not GEOMon data, but could be any data). These links must be custom-built for each individual external archive and it is therefore important to limit their number in the beginning, since a lot of programming is required in each case. The external data archives to be considered includes at least EARLINET-ASOS DB (at MPI Hamburg), AIRBASE (at MNP The Netherlands), AERONET and WDCA (at JRC Ispra). Other possible databases include the CARIBIC database, the IAGOS database, the AURA validation data centre, NDACC(etc).

Import/Export routines for NDACC data

If the data flow of NDACC relevant data are to be routed through NILU (still under discussion), it will be necessary to establish a data sharing agreement between GEOMON and the NDACC database. While CDB support the HDF formats used in NDACC, the ASCII formats used by NDACC requires developments of either EBAS or CDB in order to store these data. If the decision on data flow includes sending data directly to the NDACC database, a machine-to-machine interface between NDACC and the GEOMON data centre will instead be required.

Export routines for WDCGG

If the data flow of WDCGG (GAW, World Data Centre for Greenhouse Gases) relevant data are to be routed through NILU (still under discussion), it will be necessary to establish a data sharing agreement between GEOMON and the data centre hosted by the Japanese met office. The data formats used by GEOMON are different from those of WDCGG and it is therefore necessary to develop special export routines for these data.

General export/import routines for CDB and EBAS

Ideally, a data centre should allow users to submit data in several forms and formats, but at the same time archive the data in a unified and harmonised set of formats using a stringent metadata standard. The data centre must therefore establish data conversion routines for their data submitters. Similarly, the different users of data, may ask for data in several forms and formats and the data centre should furthermore support multiple methods for exporting data. Depending on the feedback received with the initial version of the GEOMON data centre, it may be beneficial to develop such import and export routines.

Automatic processing/checking of RD data

Some operational users such as ECMWF and other involved in GMES core services may be interested in using the Rapid Delivery data in automated procedures for model validation or assimilation. The RD data must therefore first be processed at the GEOMON data centre and provided to these users in a format they can accept. For ECMWF, this means BUFR or CREX, while other users may request XML format, netCDF, HDF or GRIB formats. In addition, there will be a need to perform simple automatic quality analyses on the data. Depending of the needs of the users of GEOMON data, an automatic processing unit for RD data will be developed. This unit will be able to check RD data and reject/accept data for further processing.

Machine-to-machine interface to GEOSS

Building on the interfaces to be developed for operational users (required for RD data), a machine-to-machine interface(s) will be developed for systems that are external to the GEOMON system. This includes the GEOSS clearinghouse that provides an overview of all registered GEOSS services and their data content, but other services (like the international version of AirNow by US EPA) may be interested in accessing the GEOMON data this way. The implementation of an interface to GEOSS will come late in the project.