WMO/IOC/UNEP/ICSU
GLOBAL CLIMATE OBSERVING SYSTEM (GCOS)

Doc. 5.1
(24.II.2009)

1st GRUAN Implementation-Coordination Meeting (ICM-1)
Norman, Oklahoma, USA
2-4 March 2009

Item 5.1

GRUAN data handling – Lead Centre outline of progress

(Submitted by GRUAN Lead Centre)

Summary and Purpose of Document

This document contains a proposal of a data handling strategy which has been developed by the GRUAN lead centre.

GRUAN Data Handling – Lead Centre outline of progress

1 Goals of GRUAN data handling

"Long-term stability" and "reference quality" are the two most important terms that describe the GRUAN goals with respect to data handling. – That implies:

- *measurements should be as accurately as possible* – The (reference) instruments will change in the next twenty years and not all sites will use the same instruments. GRUAN is and remains a heterogeneous network.
- *quality quantification (QQ)* – QA/QC of measurements is not enough. Every measured value should receive an error bar.
- *traceability* – The way measurements and data products were obtained should be traceable.

What do these facts mean for the data processing within GRUAN? A reprocessing of a complete instrument record should be possible if improvements of algorithms are developed. All steps of measuring and processing should be adequately documented.

2 Data handling policy

Figure 1 shows a proposal scheme of data flow in GRUAN. Six parts are defined. Each part is given a potential host or implementation. All parts are autonomous and can be implemented by several partners. It is possible to concentrate some parts on one host (e.g. lead centre). These parts are described in more detail in the following sections.

2.1 Collecting measuring data

There are two different data types in GRUAN and both will be collected from all sites: the normal measuring data (as raw data) from all relevant instruments and the meta data. Both types are defined in the following subsections.

How do we collect the data? A lot of options are possible, but we need one system that is available on all sites. Within GRUAN a lot of special data exists, e.g. from experimental sondes or campaigns, but also raw data and extended meta data. All this data should be collected with one system. Special services to collect data (e.g. GTS) are not available at all sites. This service is designed for near real-time dissemination (→ weather forecast). All the data is immediately globally available. However, if this option was chosen the GRUAN measurements will not have the GRUAN quality label.

The use of standard internet protocols (e.g. email, ftp, http) is another option. All sites can use it and send the data to a GRUAN data host where GRUAN QA/QC procedures are applied. Only the completely processed data (with GRUAN quality label) is then distributed to the community.

Raw data

The standard strategy in meteorological networks is to collect data which are processed and quality proven (QA/QC) directly from the site. The advantage is that this data can be disseminated to services (e.g. weather forecast) immediately. Unfortunately this procedure has a disadvantage. It is not easy to fix a problem with the data which is identified later. The applied site has to reprocess this

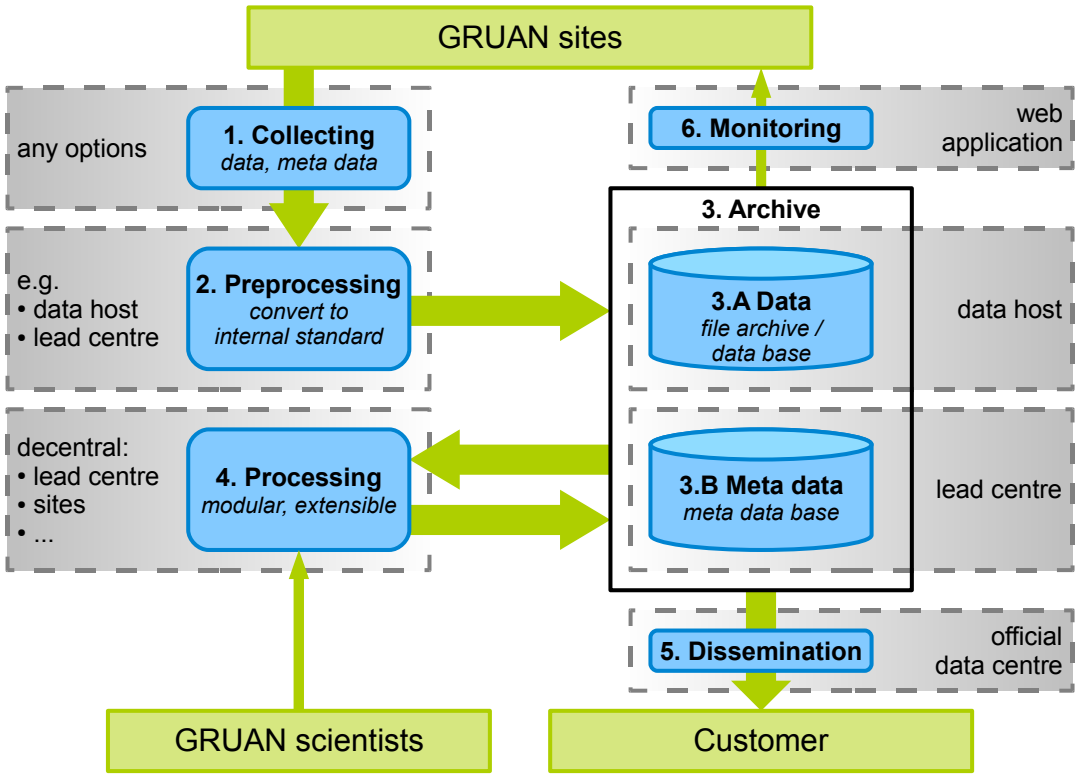

Figure 1: Proposal of data flow scheme for GRUAN

data on its own. But raw data does not always still exist.

Within GRUAN a reprocessing of complete series from all sites should be possible. This requires the archiving of raw data. Only in this case it is possible to consistently reprocess data, including the quantification of quality.

Currently, a lot of measuring systems do not properly separate between measuring and processing. In addition, some or all steps of converting the measurement signal to a target value are often directly dependent on the software version of the instrument. This software does not always have adequate documentation ("black-box" software).

For this reason we have to distinguish between two types of raw data:

- engineering raw data → which are measured, mostly electric signals
- physical raw data → which are first calculated measurands

Anyway, raw data are:

- not filtered, not corrected → Raw data should be as raw as possible.
- all data which are needed to calculate the target variables and to quantify their quality are recorded

Meta data

Meta data is additional information which is collected together with the normal measuring data. It describes the measurement system, the date and the location. It also gives information about the in-

strument and the meteorological conditions.

For example, for a radiosonde ascent meta data are:

- basic facts: when, what, where, (who)
- how (assembly of rig): balloon, parachute, string length, instrument position on rig
- meteorological parameter on the ground: pressure, temperature, relative humidity, cloud information
- ground check data, coefficients, ...

This meta data should contain all additional information to categorise and to understand the target variables.

2.2 Preprocessing

In this strategy preprocessing is the step to test and to harmonise the data. The collected data consisting of raw data and meta data is imported and tested for completeness and consistency. This is the first step of the GRUAN-internal QA/QC.

Now, the data is converted into the standard GRUAN data format and the files are saved in the data archive. It is also useful to save the original files as backup. The meta data is analysed and stored in the "meta data database".

After preprocessing, all information about a individual measurement can be obtained from the meta database. The conversion to a standard data format at that stage seems useful since we do not have to convert any data in the following steps.

The standard data format and the additional information in the "meta data database" allows for a simple interface for the data access and will greatly facilitate the work of the different partners who actually process the data. These partners will have access to the data archive and the meta database.

Data format

Within GRUAN different types of instruments and sources are used, like different radiosondes, GPS receivers, ceilometers, lidars, etc. Each of these different sources require a data format that allows to store its specific data. The data format should be readable over a long time and should have an open standard. Different groups of users will use this data and therefore it should be easy to use and self-describing. A data format with existing and free (open-source) software libraries for reading and writing is the best solution.

We suggest as the internal GRUAN data format the NetCDF format (Network Common Data Form) because it fulfils all the necessary requirements.

2.3 Archive

The GRUAN archive consists of two independent parts – a file archive and a meta data base.

A) The file archive contains all GRUAN data as consistently named files (backup of original data, converted raw data, processed data).

B) The meta database holds all information which could be relevant to the use of GRUAN data:

- information on the sites: location, host institution, measurement systems, instrumentation, …
- the measurement itself – see chapter 2.1 / Meta data
- the processed data products – including level and versions, used algorithms and software for processing, …

Both parts of the archive possess a defined interface (e.g. web service) that gives access to the data and meta data. This access is limited to the GRUAN community.

2.4 Processing

Processing is the central part of data handling. Here all processing steps which are needed to produce the data products (level 2 + 3) from the raw data (level 1) are included. A modular structure of the processing scheme allows flexibility which is needed for our heterogeneous network. The data and the meta data can be exchanged with the archive over the defined interface. Figure 2 shows this schematically.

Within the processing any modules (software applications) can be used, e.g. for testing, filtering, calculation, correction, interpolation, quality quantification, etc. For every specific data product a special processing procedure is defined which includes the used modules, their order and parametrisation. These procedures are recorded in the meta database (3.B) which leads to an excellent traceability.

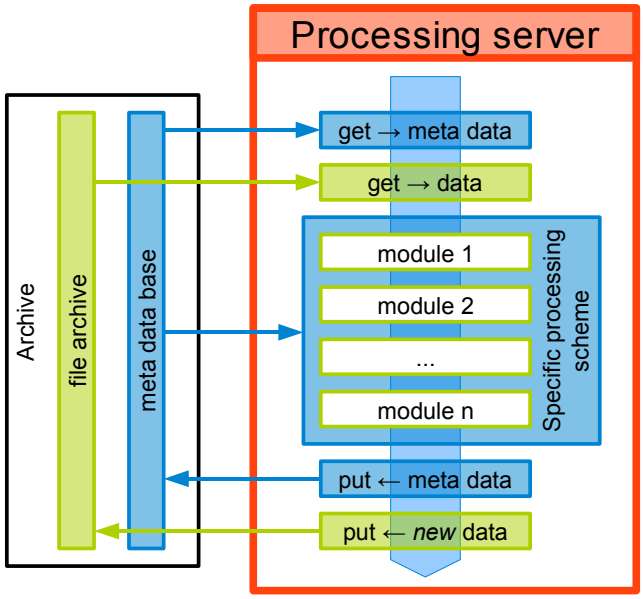

Figure 2: Processing scheme

Each defined procedure of the processing can be implemented separately. Therefore the data processing can be split to different hosts (processing servers) at different locations. This allows GRUAN scientist to develop improved or new algorithms/procedures that can easily be plugged in the processing scheme. New processing software can be used immediately in a consistent way for the data of all stations without additional effort. This approach promotes team-working within GRUAN, facilitates the comparison different methods and therefore advances the quality of GRUAN data products.

Processing software

All data products of GRUAN including QQ should be traceable and verifiable. For this reason, traceability and validation also apply for the processing with all its components (software applications).

The processing software used in GRUAN should have the following properties:

- easily extendible (modular design with a open interface)
- complete documentation
- version control (e.g. SVN)
- free access and free use

The GRUAN lead centre publishes all software developed by its own as open source.

Level of data "products"

Data and data products can be classified into 4 levels:

0 original files (as backup)

1 raw data (after preprocessing)

2 processed data including error bar for each single value, no use of independent measurements

3 "best possible" profile → composite of independent measurements

2.5 Dissemination of data products

The dissemination of processed GRUAN data (data products) is realized with an established data centre (e.g. NCDC).

The website of GRUAN contains all information about the data products. It offers:

- a possibility to search data products
- all relevant documentation
- special software for easy use of data (like a viewer)

2.6 Monitoring

The monitoring serves to check the status of current data flow and network (collection → processing → dissemination). A monitoring tool should be accessible from all sites and all partners. It can be implemented as a web application. All needed status information is situated in the meta database (3.B).

3 Discussion

The presented strategy is a proposal of the GRUAN lead centre. It is open to discussion. Most importantly, the following questions need to be resolved:

- How do we collect data and meta data within the GRUAN network?

- Is the collection of raw data feasible?
- How do we handle "black-box" software with respect to quality quantification and error handling?
- How promptly should the data products be distributed?