Session 5
Data dissemination and data policy

Outline of progress to date

Michael Sommer

GRUAN Lead Centre, DWD

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

Contents

→ Goals

→ Strategy of data handling

      1. Collection & 2. Preprocessing

      3. Archive

      4. Processing
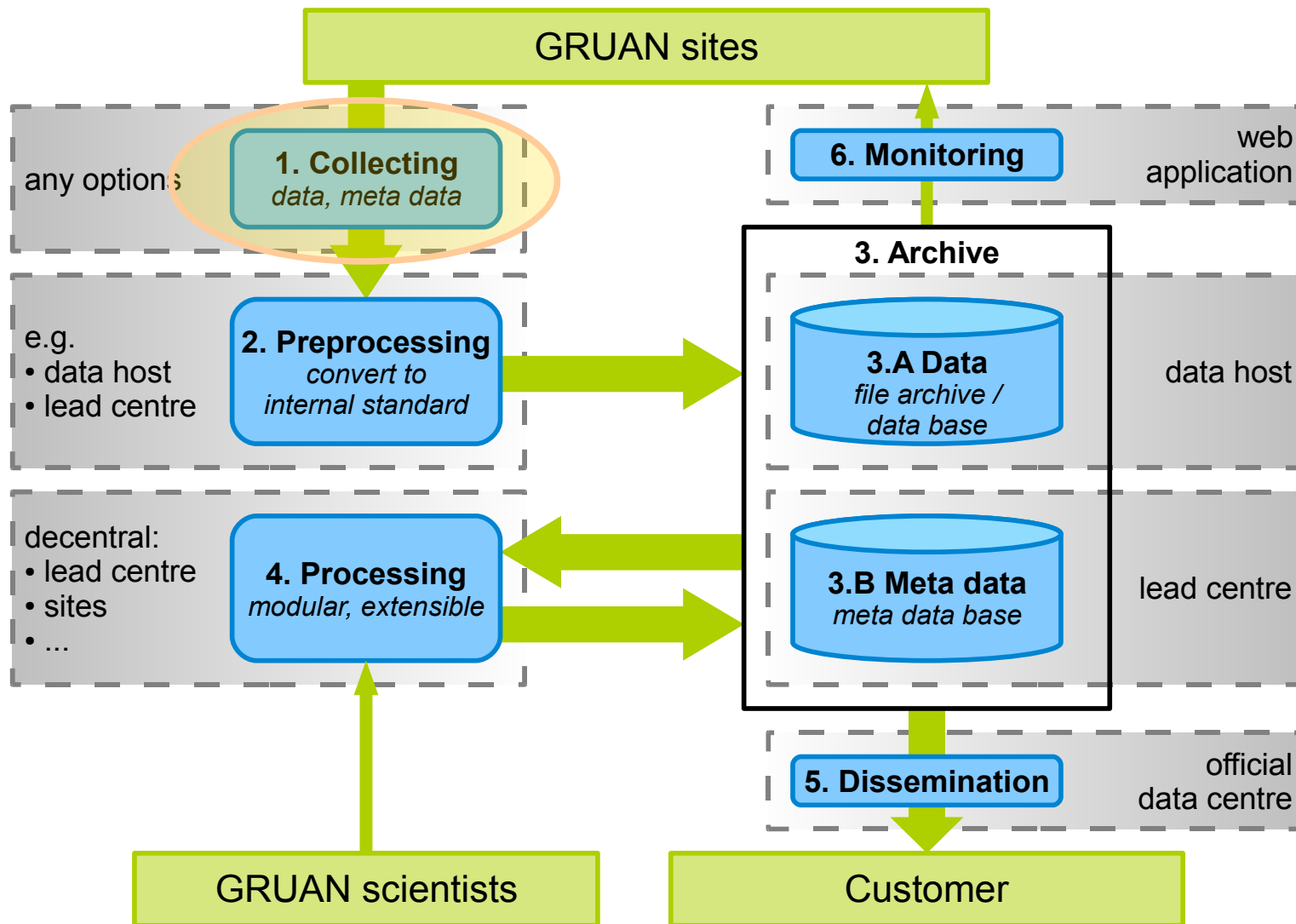
      5. Dissemination & 6. Monitoring

→ Discussion

Goals of GRUAN data handling

Long-term stability & reference quality imply:

➜ as accurately as possible → heterogeneous measuring (instruments/network)

➜ quality quantification → error bar for all measured values

➜ traceability → still in 20 years !!!

What do these facts mean for the data processing within GRUAN?

➜ **reprocessable** → improvements of algorithms should be of use for complete series

➜ **traceable** → all steps of measuring and processing should be adequately documented (meta data)

1. Collecting measuring data

➔ What do we collect?

- any measuring data (*raw data*) which is relevant for GRUAN
  (first of all → priority one instruments)

- *meta data* of measurements
  (e.g. conditions, equipment, used software, operator, problems, ...)

➔ How do we collect?

- a lot of variants are possible

  - *elaborate* → special services to collect data e.g. GTS

  - *simple* → email, ftp, http

- **agreement is necessary**
  → optimal integration into the existent data flow of sites

Raw data & meta data

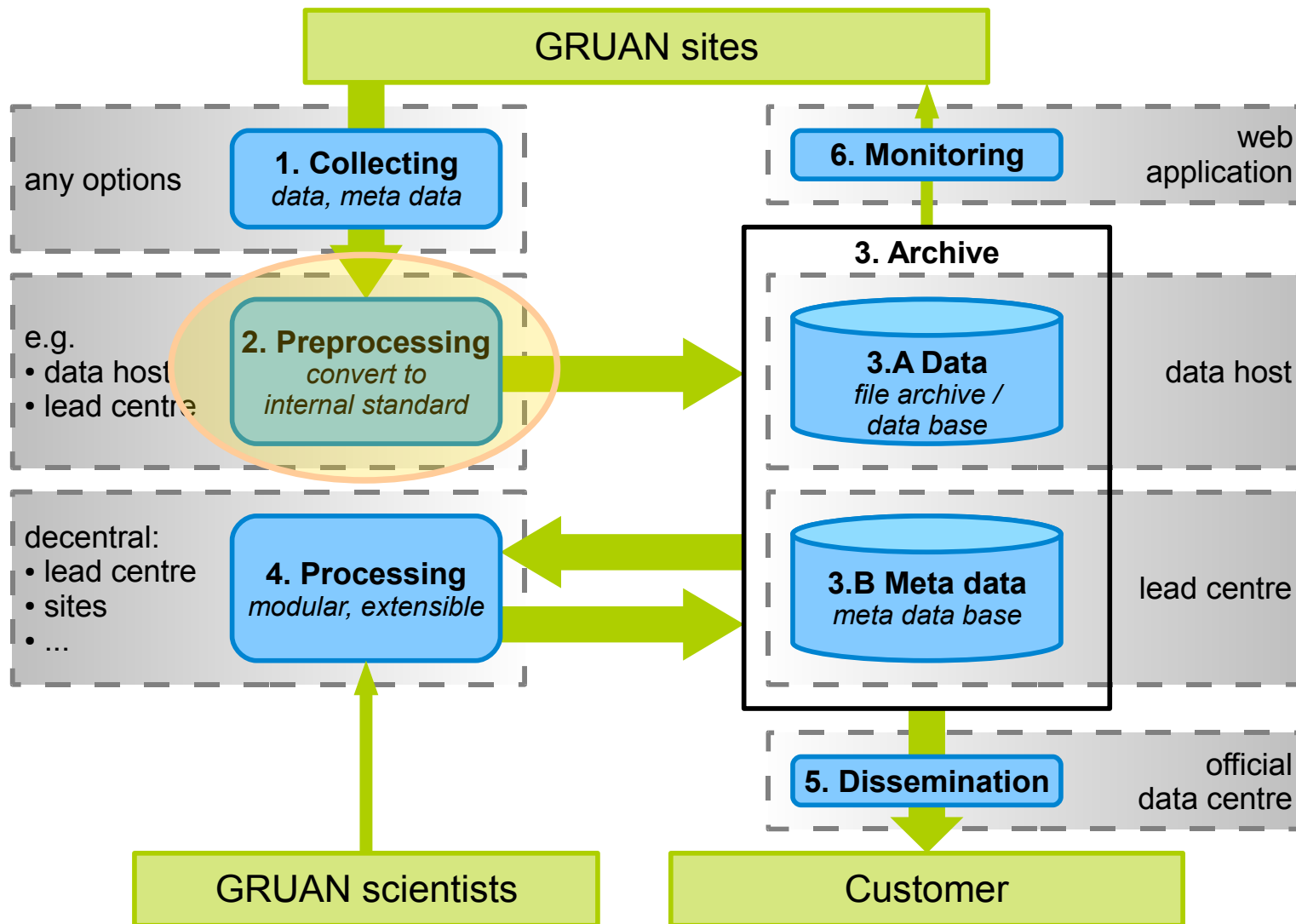| What are raw data in GRUAN?	What are meta data in GRUAN?
A) Engineering raw data measured signals *(frequencies, voltage, ...)*	**Additional information**
B) Physical raw data first calculated measures *(pressure, temperature, …)*	**Example – a radiosonde ascent** • basic facts: when, what, where, (who) • how (assembly of rig): balloon, parachute, string length, position • meteorological parameters • ground check, coefficients, …
Summarised • raw data are not filtered, not corrected, … • all data which are needed to calculate the target variables and to quantify their quality	**Summarised** Meta data are all additional information to categorise and to understand the target variables.

GRUAN sites

1. Collecting
data, meta data

any options

6. Monitoring

web application

2. Preprocessing
convert to internal standard

e.g.
• data host
• lead centre

3. Archive

3.A Data
file archive / data base

data host

4. Processing
modular, extensible

decentral:
• lead centre
• sites
• ...

3.B Meta data
meta data base

lead centre

5. Dissemination

official data centre

GRUAN scientists

Customer

2. Preprocessing

→ Import all collected *raw data* and *meta data*

→ Test the integrity of data

→ Convert to **the** standard GRUAN data format

→ Memory converted files in data archive
  (+ original files as backup)

→ Inform "meta data database" of results from preprocessing

Data format

Which requirements for the data format exist?

➜ useable for *all imaginable* GRUAN data

- radiosonde, GPS receivers, ceilometer, lidar, …

➜ open standard

➜ easy to use

- existent and free software (libraries) for reading and writing
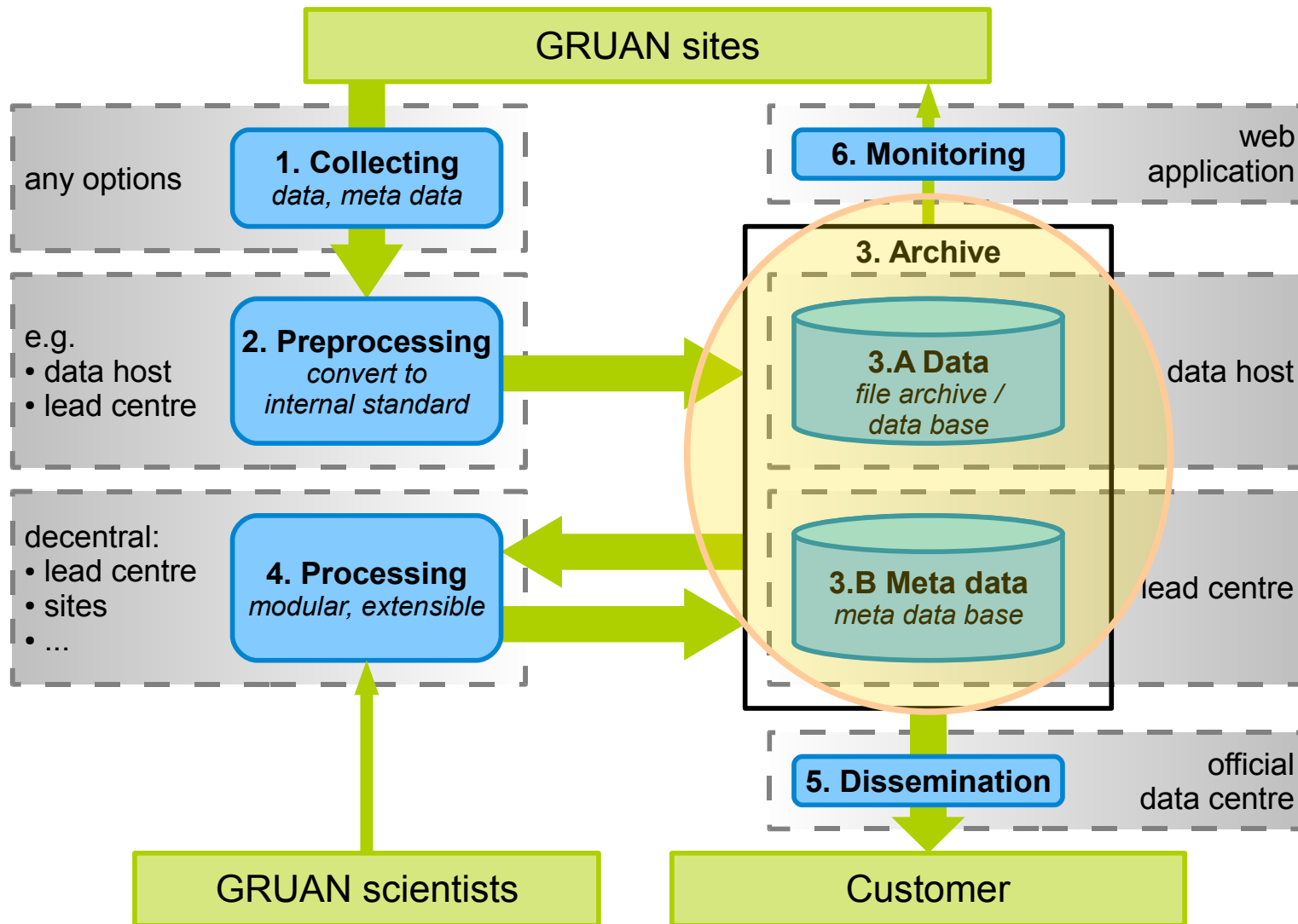- self-describing

Proposal for data format

NetCDF – Network Common Data Form (Unidata / UCAR)

Pros
• open standard
• self-describing
• free software libraries for reading and writing NetCDF files (open source)
• independent from platform and computer language
• popular

Cons
• Not human-readable

GRUAN sites

1. Collecting
data, meta data

any options

2. Preprocessing
convert to internal standard

e.g.
• data host
• lead centre

4. Processing
modular, extensible

decentral:
• lead centre
• sites
• ...

6. Monitoring

web application

3. Archive

3.A Data
file archive / data base

data host

3.B Meta data
meta data base

lead centre

5. Dissemination

official data centre

GRUAN scientists

Customer

3. Archive

GRUAN Archive

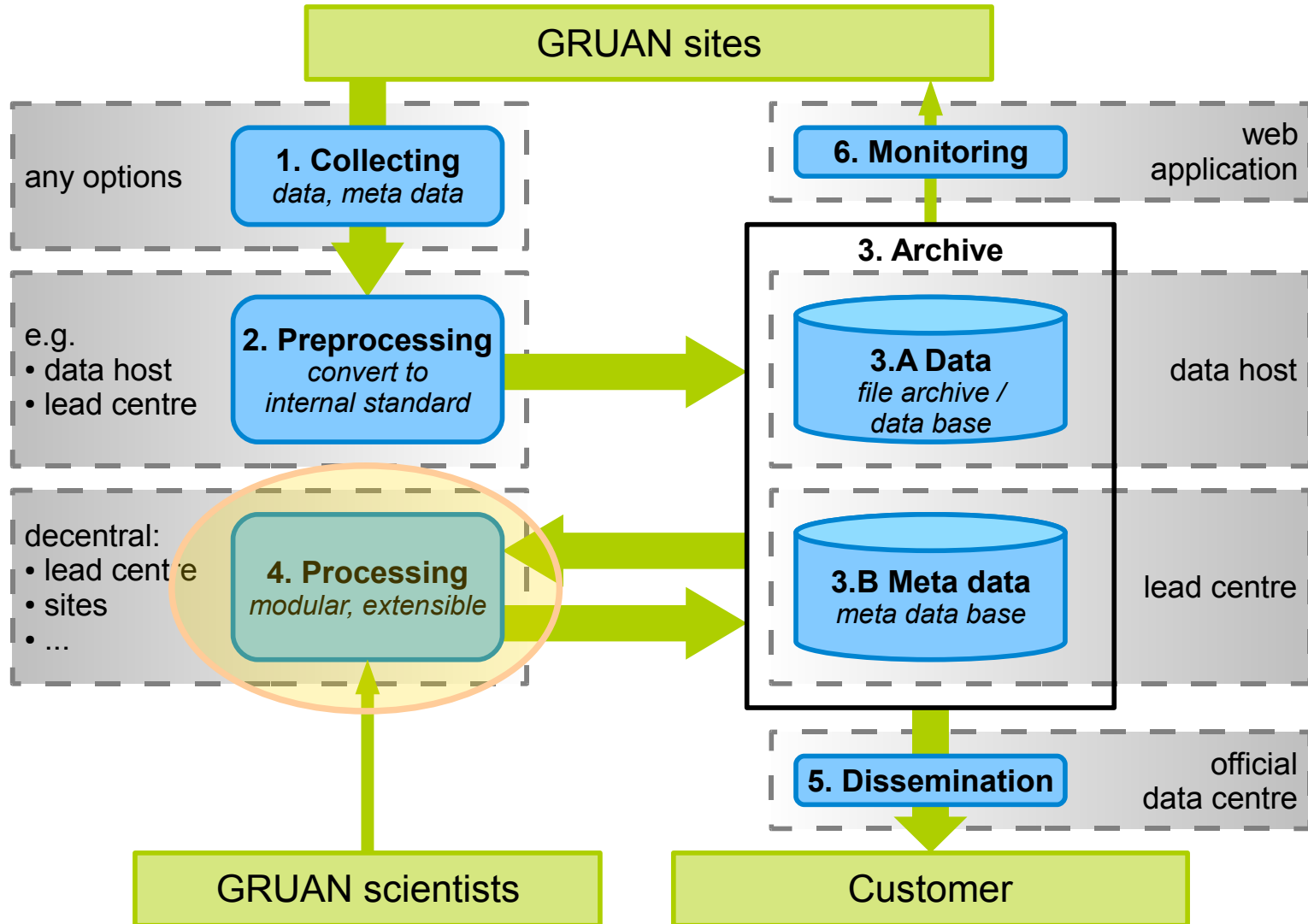### A) Data
file archive / database

Memory of:
- original data files
  - ➔ as backup additional to sites
- converted standardised files
  - ➔ from preprocessing
- processed data / data products

B) Meta data
meta data base

Information about:
- sites
  - ➔ location, measurement systems, instruments, …
- measurements
  - ➔ all "relevant" infos
- processing
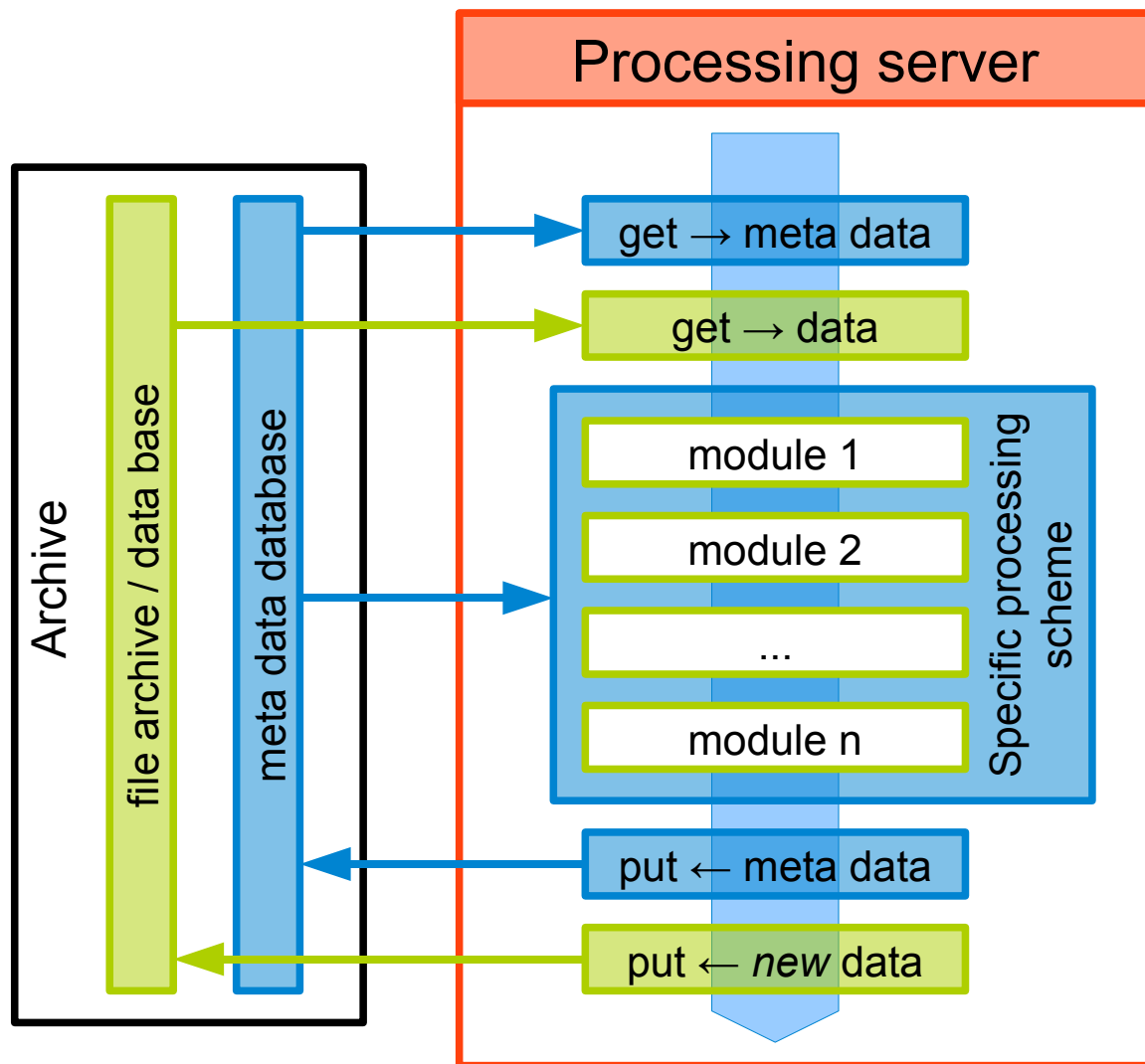 - ➔ level, versions, used algorithm, software version, ...

4. Processing

Modular layout

→ standard interface for archive communication

→ specific modules to:

test, filter, calculation, error handling (QA, QQ), correction, interpolation, merging, etc.

Traceability

→ all processing infos in meta data base

→ definition of specific processing schemes

Processing server

- get → meta data
- get → data
  - module 1
  - module 2
  - ...
  - module n
- Specific processing scheme
- put ← meta data
- put ← *new* data

Archive
- file archive / data base
- meta data database

Processing software

What do we need?

- complete **traceable** processing (incl. quality quantification)
- **verifiable** → *by each customer, who would like this*

Proposal for processing software → **open to discussion**

- extendible (modular design with a open interface)
- complete documentation
- version control
- free access and free use

> Developed software from Lead Centre are
> **Open Source Software**

Level of data "products"

0   **Original data files**

- source for preprocessing
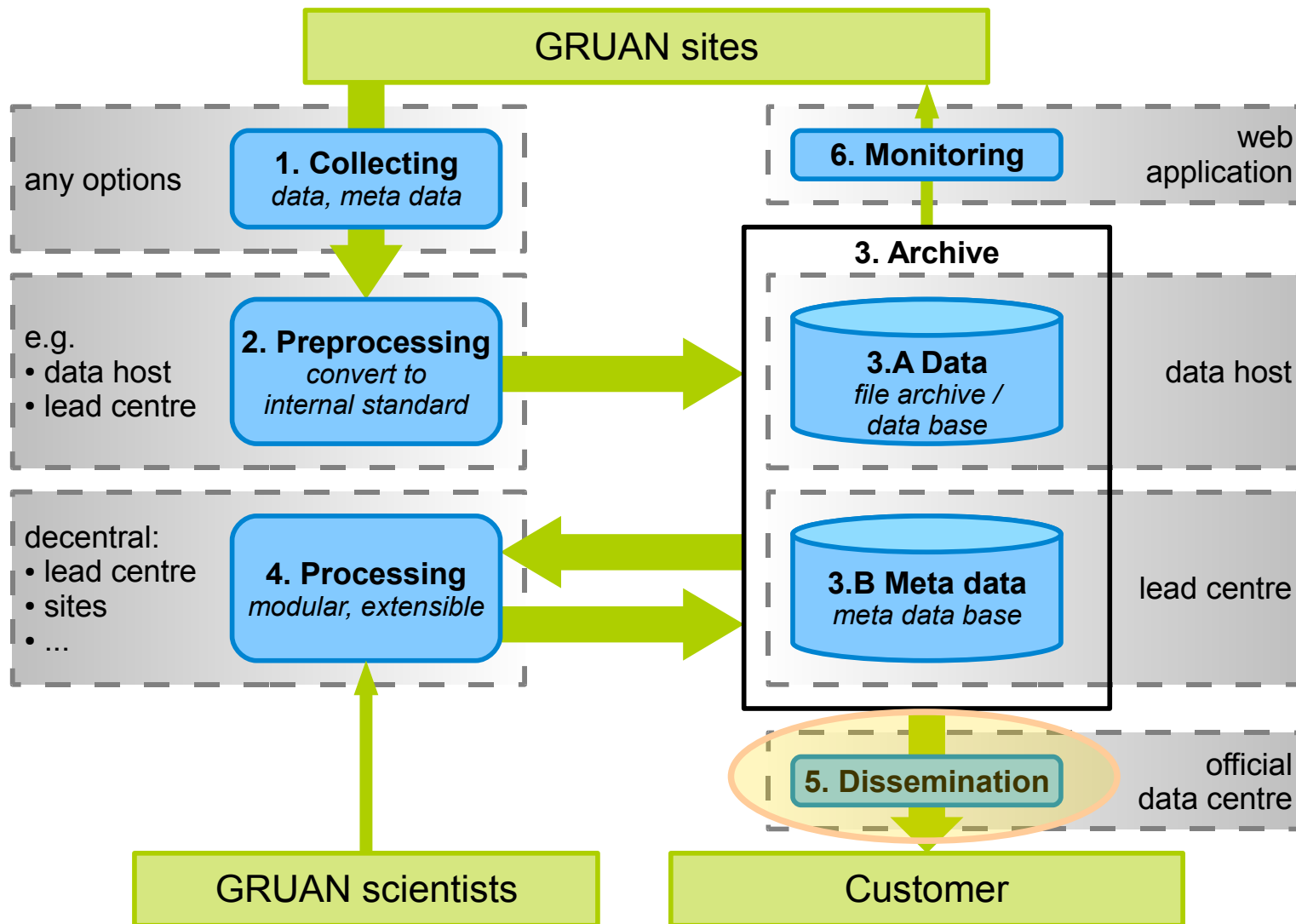- backup

---

1   **Raw data**

- *see → definition of raw data*

2   **Processed measuring data** + error bar on all values

- e.g. interpolated, filtered, corrected, …
- **no use** of independent observations

3   **"Best possible" profile** – composite data (merging)

- **use** of independent observations
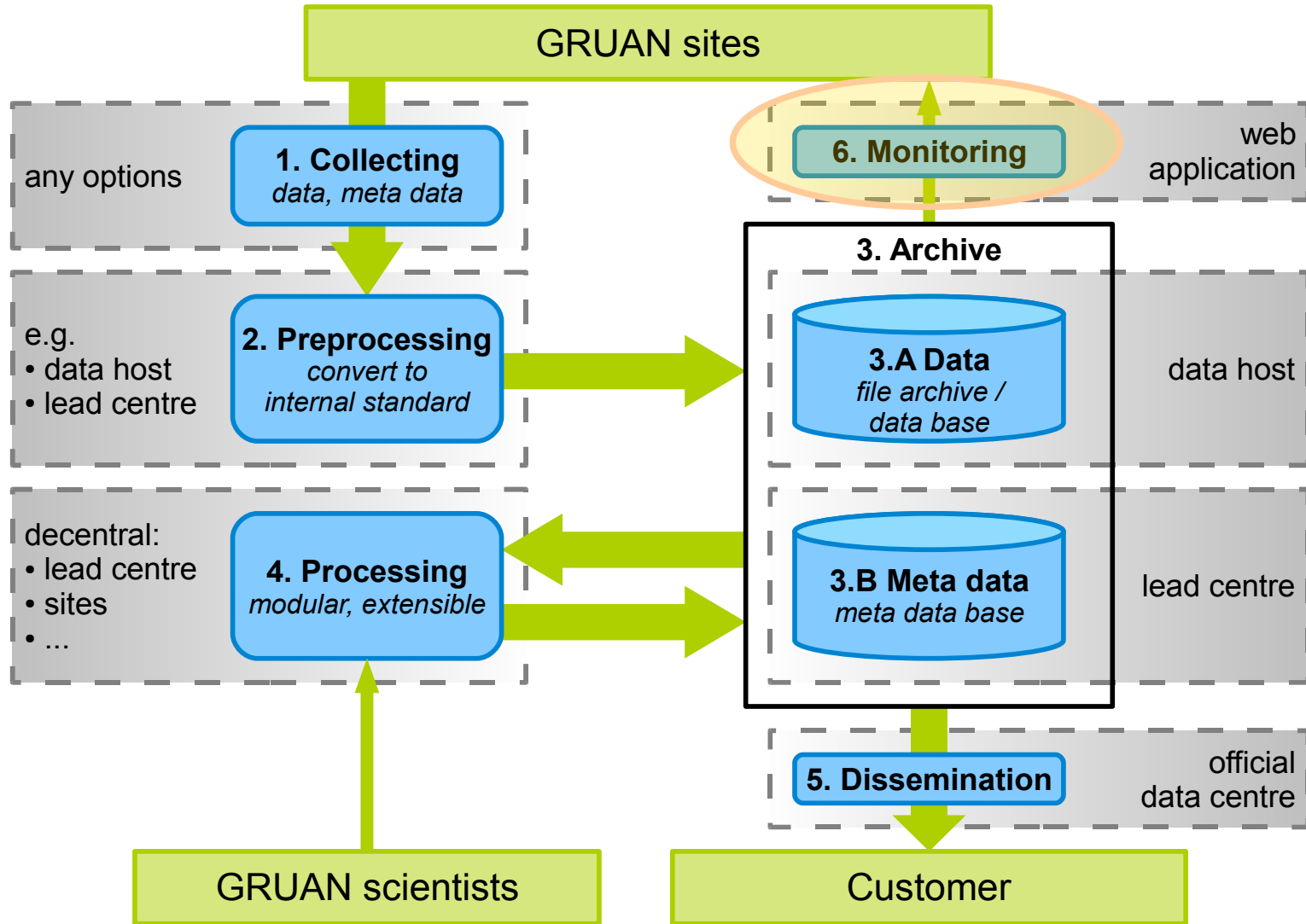
5. Dissemination of data products

Objectives of dissemination:

- integration into the generally usual distribution ways
- use of an existent data centre (e.g. NCDC)

How promptly should the data products be distributed?

On GRUAN web site:
- search of data products
- documentation
  - → *all information about measurements and data products*
- special software for easy usage of data (like a viewer)

6. Monitoring

➜ What does it serve for?

- status of network
- status of processing
- detection of problems

➜ Who has access to the monitoring?

- all sites
- Lead Centre
- science partner

Implementation as a web application

Discussion

➔ How do we collect data and meta data?

➔ Is the collection of the *raw data* feasible?

➔ How do we handle "black-box" software?
→ *quality quantification and error handling*

➔ How promptly should the data products be distributed?

➔ How could you contribute?
→ *resources, existent software solutions, processing methods*