EDG Final Review Demonstration

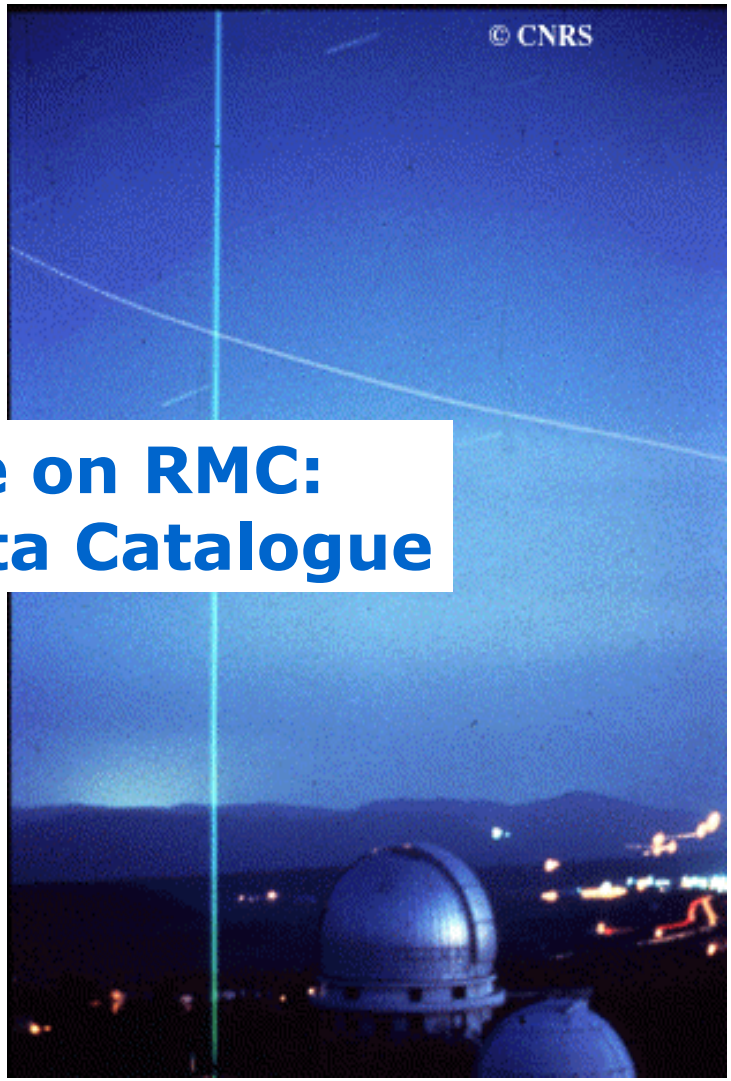
WP9 Earth Observation Applications

Meta data usage in EDG

Christine Leroy, Wim Som de Cerff
leroy@ipsl.jussieu.fr, sdecerff@knmi.nl

DataGrid is a project funded by the European Commission under contract IST-2000-25182

Earth observation Meta data usage in EDG



Focus will be on RMC: Replica Metadata Catalogue

- Validation usecase: Ozone profile validation
- Common EO problem: measurement validation
- Applies to (almost) all instruments and data products, not only GOME, not only ozone profiles
- Validation consists of finding, for example, less than 10 profiles out of 28,000 in coincidence with one lidar profile for a given day
- Tools available for metadata on the Grid: RMC, Spitfire

Demonstation outline

Replica Metadata Catalogue (RMC) usage

1) Profile processing
   Using RMC to register metadata of resulting output

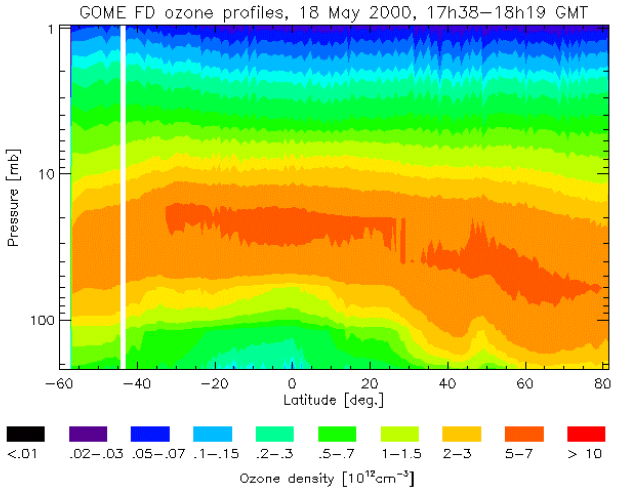
2) Profile validation
   Using RMC to find coincidence files

3) RMC usage using a web interface
   Will show the content of the RMC, the attributes we use.

4) Show result of the validation

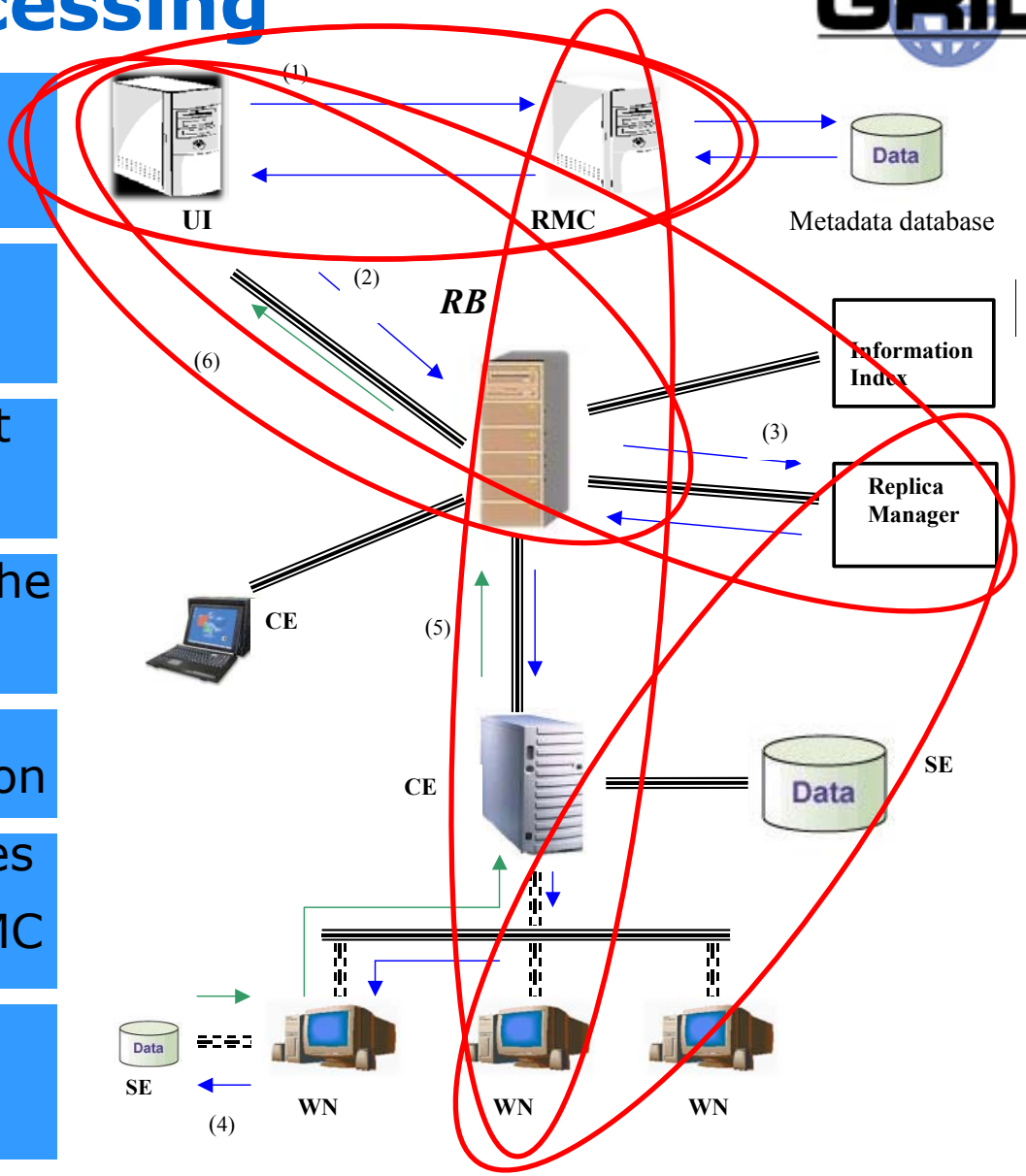
File type	No. of entries in RMC
Level 1	3,402
NNO profiles	6,092
Opera profiles	14,745
Lidar files	645
Total:	32,514*

(* including test and aux. Data)



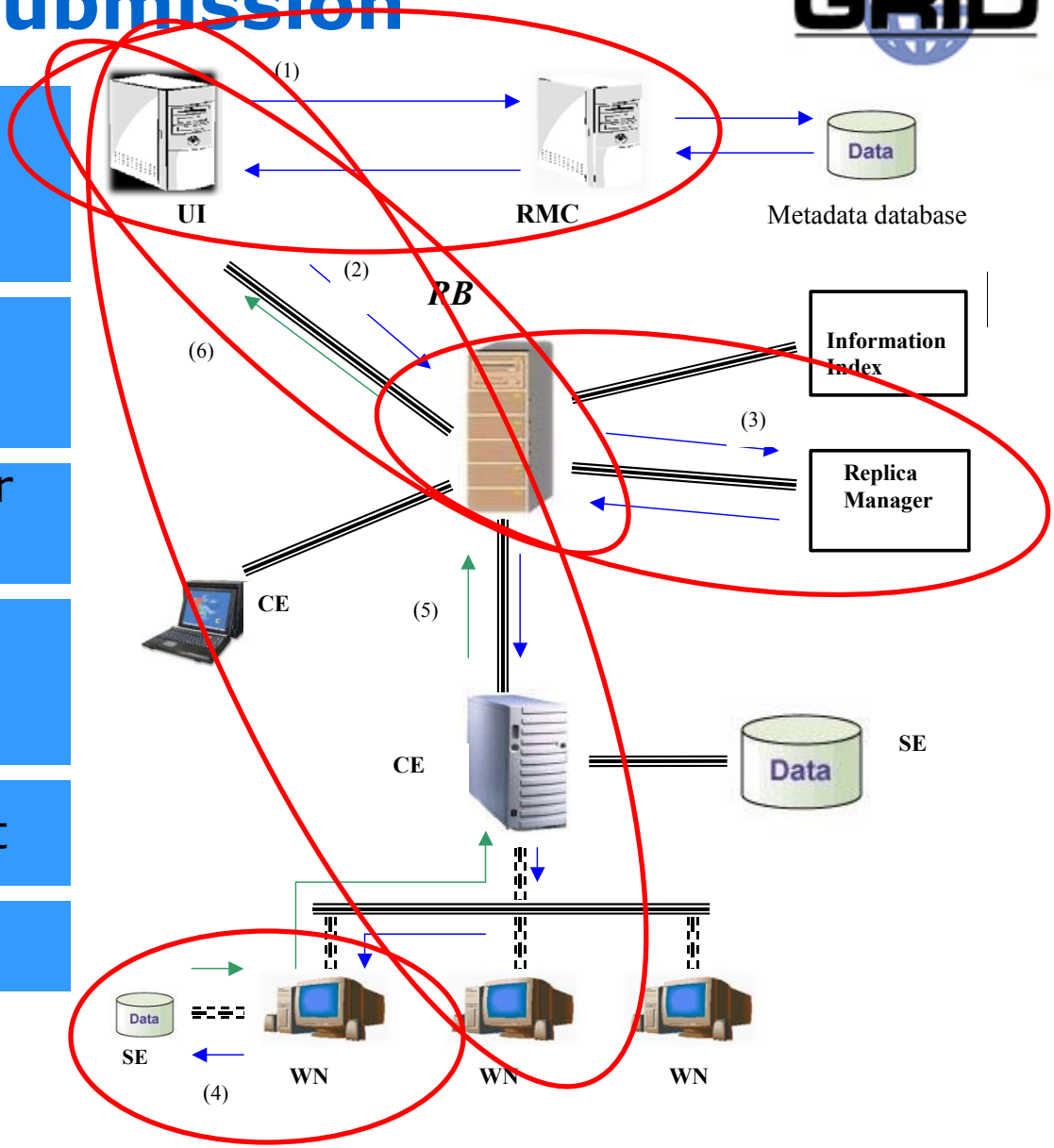
GOME NNO Processing

1. select a LFN from precompiled list of non-processed orbits
2. verify that the Level1 product is replicated on some SE
3. verify the Level2 product has not yet been processed
4. create a file containing the LFN of the Level1 file to be processed
5. create a JDL file, submit the job, monitor execution
6. During processing profiles are registered in RM and metadata is stored in RMC
7. query the RMC for the resulting attributes

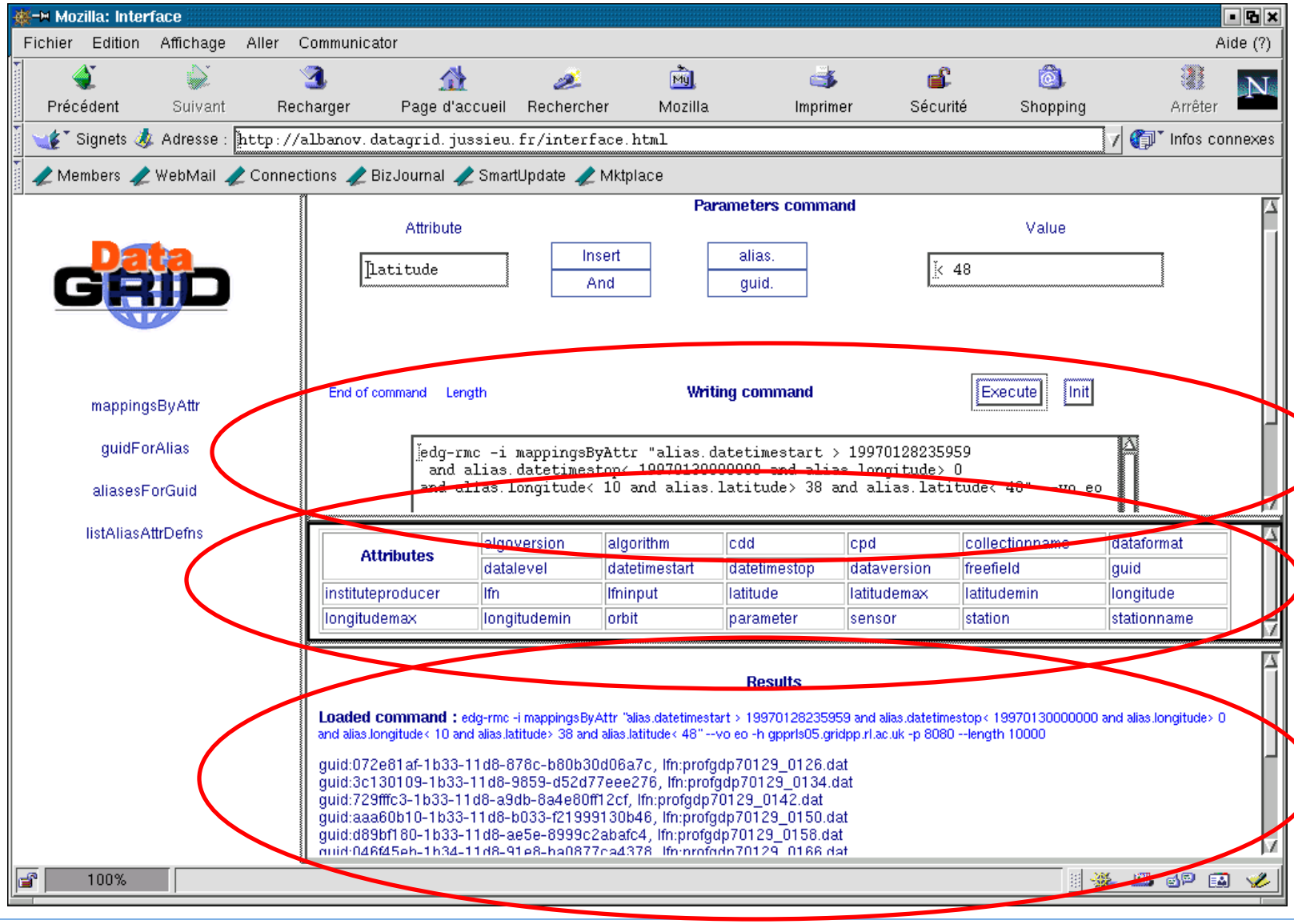


Validation Job submission

1. Query RMC for coincidence data LFNs (Lidar and profile data)
2. Submit job, specifying the LFNs found
3. Get the data location for the LFNs from RM
4. Get the data to the WN from the SE and start calculation
5. Get the output data plot
6. Show the result



RMC usage: attributes



Command area

All attributes Of WP9 RMC

Result area

Metadata tools comparisons

Replica Metadata Catalogue Conclusions, future direction:

- RMC provides possibilities for metadata storage
- Easy to use (CLI and API)
- No additional installation of S/W for user
- RMC performance (response time) is sufficient for EO application usage
- More database functionalities are needed: more data types, polygon queries, multiple tables, restricted access (VO, group, sub-group)

Many thanks to WP2 for helping us preparing the demo

Backup slides

EO Metadata usage

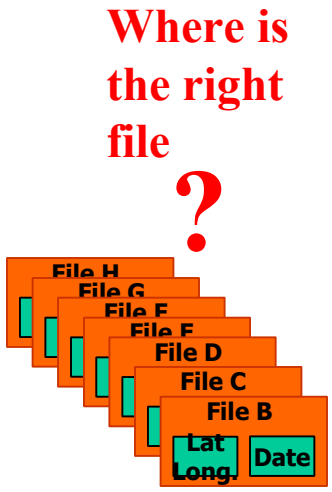
Questions adressed by EO Users:
How to access metadata catalogue using EDG Grid tools?

Context:

- In EO applications, large number of files (millions) with relative small volume.
- How to select data corresponding to given geographical and temporal coordinates?
- Currently, Metadata catalogues are built and queried to find the corresponding files.

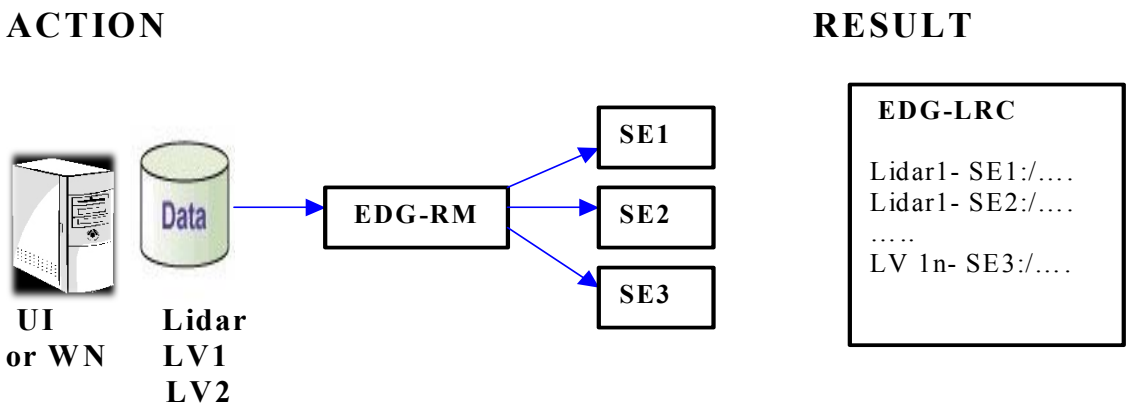
Gome Ozone profile validation Usecase:

- ~28,000 Ozone profiles/day or 14 orbits with 2000 profiles
- Validation with Lidar data from 7 stations worldwide distributed
- Tools available for metadata on the Grid: **RMC**, **Spitfire,** **Muis** (operational ESA catalogue) via the EO portal



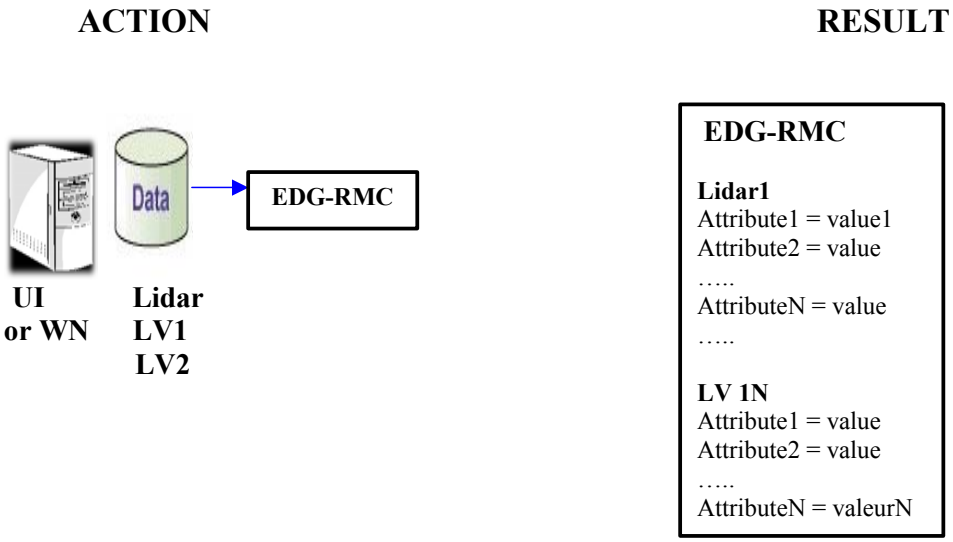
Data and Metadata storage

- Data are stored on the SEs, registered using the RM commands:



- Metadata are stored in the RMC, using the RMC commands

Link RM and RMC:
Grid Unique Identifier (GUID)



Usecase: Ozone profile validation

Step 1:	Transfer **Level1 and LIDAR** data to the Grid Storage Element
Step 2:	Register **Level1** data with the Replica Manager Replicate to other SEs if necessary
Step 3:	Submit jobs to process **Level1** data, produce Level2 data
Step 4:	Extract metadata from level 2 data, store it in database using Spitfire, store it in Replica Metadata Catalogue
Step 5:	Transfer **Level2** data products to the Storage Element Register data products with the Replica Manager
Step 6:	Retrieve coincident level 2 data by querying Spitfire database or the Replica Metadata Catalogue
Step 7:	Submit jobs to produce Level-2 / LIDAR **Coincident** data perform **VALIDATION**
Step 8:	Visualize Results

Which metadata tools in EDG?

Spitfire

- Grid enabled middleware service for access to relational databases.
- Supports GSI and VOMS security
- Consists of:
  - the **Spitfire Server** module
    Used to make your database accessible using Tomcat webserver and Java Servlets
  - the **Spitfire Client** libraries
    Used from the Grid to access your database (in Java and C++)

Replica Metadata Catalogue:

- Integral part of the data management services
- Accessible via CLI and API (C++)
- No database management necessary

Both methods are developed by WP2

Focus will be on RMC

Scalability (Demo)

- this demonstrates just one job being submitted and just one orbit is being processed in a very short time
- but the application tools we have developed (e.g. **batch** and **run** scripts) can fully exploit possibilities for parallelism
- they allow to submit and monitor tens or hundreds of jobs in one go
- each job may process tens or hundreds of orbits
- just by adding more LFNs to the list of orbits to be processed
- **batch –b** option specifies the number of orbits / job
- **batch –c** option specifies the number of jobs to generate
- used in this way the Grid allows us to process and register several years of data very quickly
- example: just 47 jobs are needed to process 1 year of data (~4,700 orbits) at 100 orbits per job
- this is very useful when re-processing large historical datasets, for testing differently 'tuned' versions of the same algorithm
- the developed framework can be very easily reused for any kind of job

GOME NNO Processing – Steps 1-2

Step 1) select a LFN from precompiled list of non-processed orbits

```
>head proclist
70104001
70104102
70104184
70105044
70109192
70206062
70220021
70223022
70226040
70227033
```

Step 2) verify that the Level1 product is replicated on some SE

```
>edg-rm --vo=eo lr lfn: 70104001.lv1

srm://gw35.hep.ph.ic.ac.uk/eo/generated/2003/11/20/file8ab6f428-1b57-11d8-
    b587-e6397029ff70
```

GOME NNO Processing – Steps 3-5

Step 3) verify the Level2 product has not yet been processed

```
>edg-rm --vo=eo lr lfn: 70104001.utv

Lfn does not exist : lfn:70104001.utv
```

Step 4) create a file containing the LFN of the Level1 file to be processed

```
>echo 70104001.lv1 > lfn
```

Step 5) create a JDL file for the job

(the **batch** script outputs the command to be executed)

```
>./batch nno-edg/nno -d jobs -l lfn -t

run jobs/0001/nno.jdl -t
```

GOME NNO Processing – Steps 6-7

Step 6) run the command to submit the job, monitor execution and retrieve results

```
>run jobs/0001/nno.jdl –t
Jan 14 16:28:45 https://boszwijn.nikhef.nl:9000/o1EABxUCrxzthayDTKP4_g
Jan 14 15:31:42 Running grid001.pd.infn.it:2119/jobmanager-pbs-long
Jan 14 15:57:36 Done (Success) Job terminated successfully
Jan 14 16:24:01 Cleared user retrieved output sandbox
```

Step 7) query the RMC for the resulting attributes

```
./listAttr 70517153.utv
lfn=70517153.utv
instituteproducer=ESA
algorithm=NNO
datalevel=2
sensor=GOME
orbit=10844
datetimestart=1.9970499E13
datetimestop=1.9970499E13
latitudemax=89.756
latitudemin=-76.5166
longitudemax=354.461
longitudemin=0.1884
```