

# Data Manager

how to deal with data  
coming from experiments ?

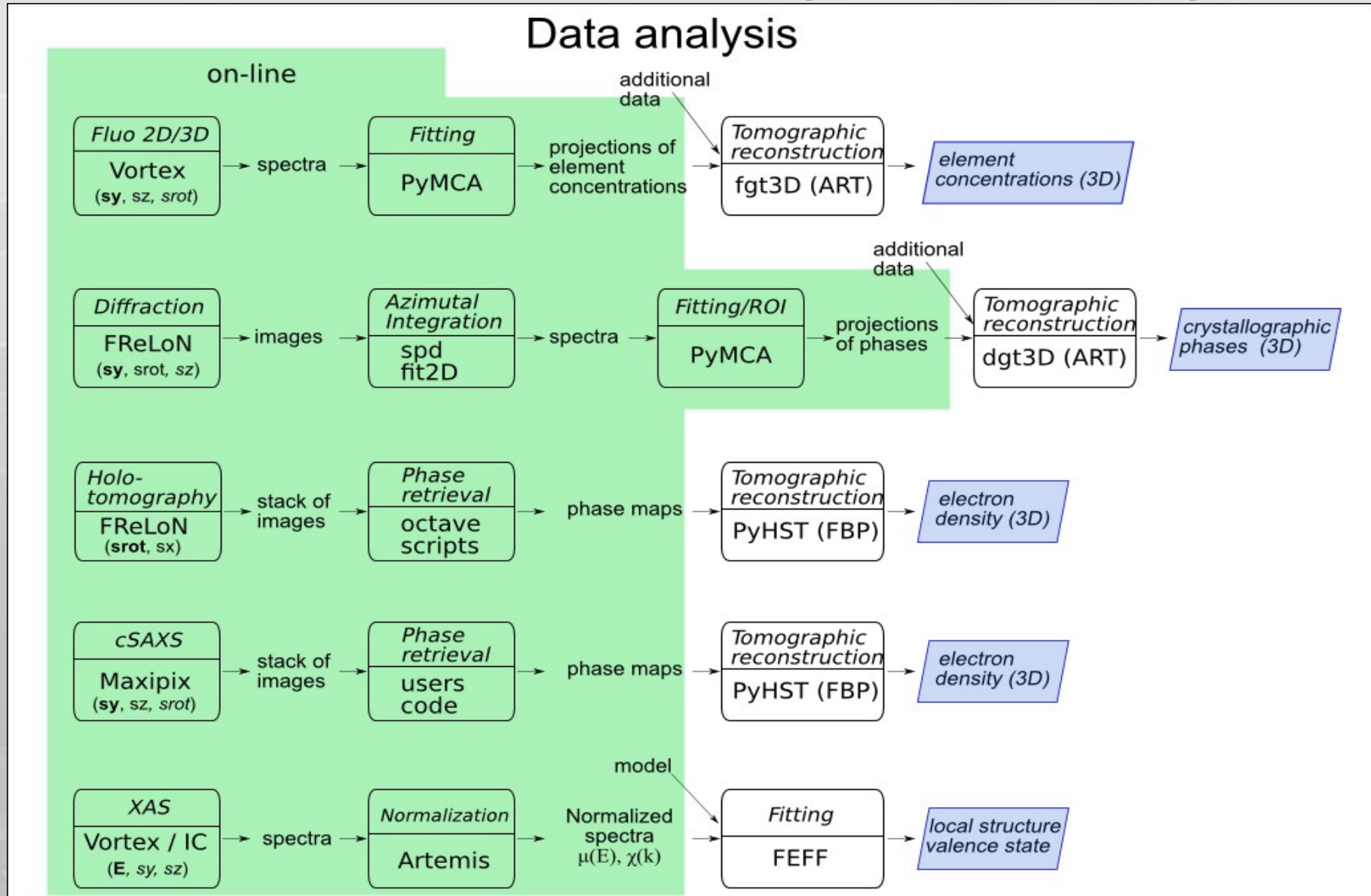
*Cyril GUILLOUD - ESRF - BLISS team*

*WP 10 Workshop  
Friday 14th  
January  
2011*

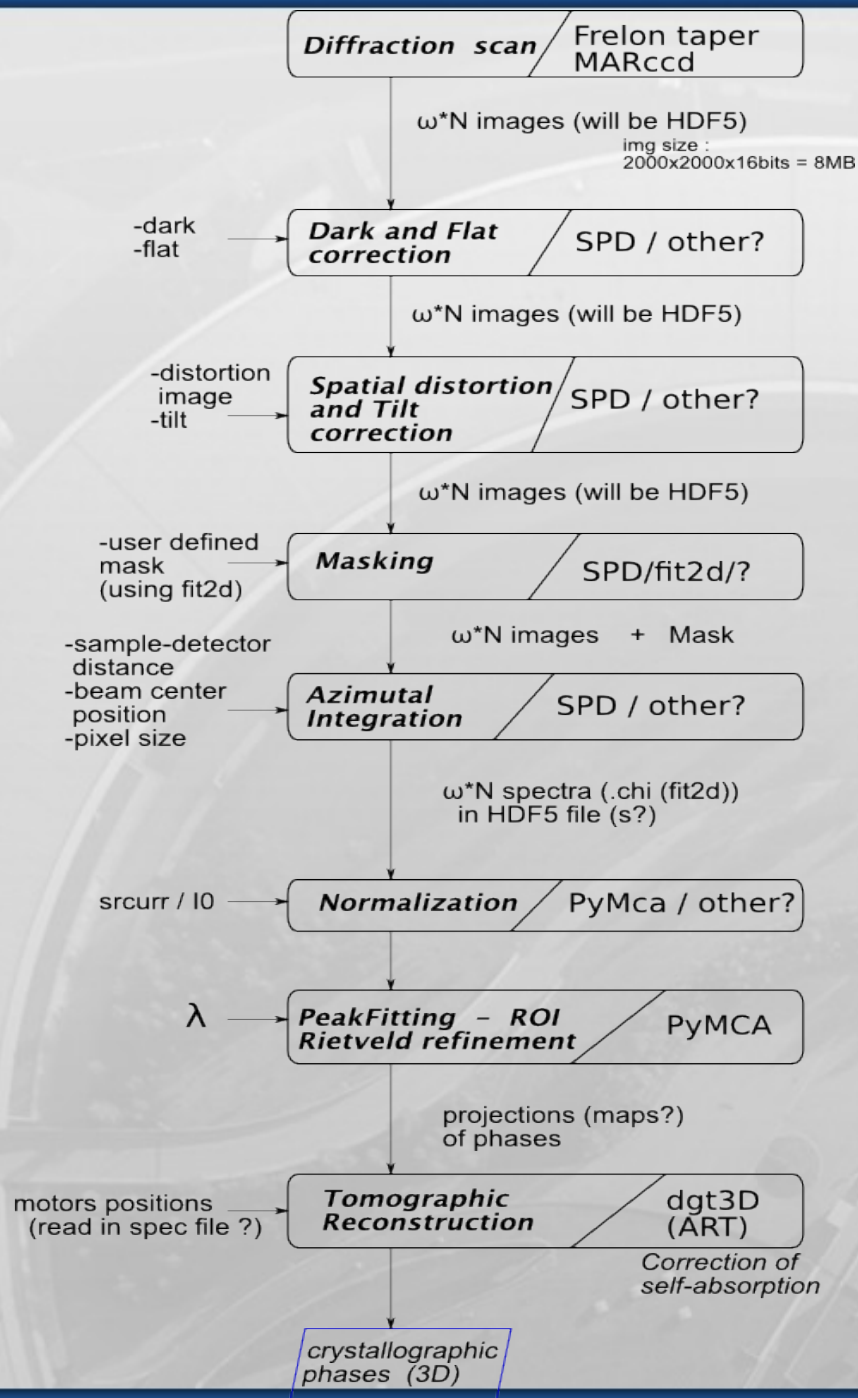
# Origin

- diffraction tomography experiment on ID22
- Involves a lot of tools for data analysis
- Tentative to use EDNA to wrap all tools

# ID22 dataflows (simplified)



# Diffraction tomography dataflow



# Today's situation at ESRF : data behavior

- The sequencer drives in the same code :
  - The experiment control (ex : motor motion)
  - The data acquisition (ex : ccd reading)
  - The data saving (ex : spec or .edf file writing)
- Data are saved at various levels :
  - SPEC built-in
  - SPEC macros
  - Device Servers
  - Treatment tools

# Today's situation at ESRF : data formats

- EDF : (ESRF Data Format)
  - Text for headers + binary data
  - Mainly used for images and MCA
  - PyMca Oxidis Onze ImageJ readable
- SPEC file : (text based)
  - Newplot, PyMca and other tools readable
  - No storage of treated data.
  - No formal semantic

# Today's situation at ESRF : data analysis

- “Analysis workflow” is a patchwork of different pieces of code more or less decent.
- Many origins:
  - Bliss ; SciSoft ; Scientists ; External tools ; External users
- Many languages :
  - C; C++; python; SPEC; Matlab; java; Octave ...
- Variable complexity
  - From text arrangement to tomo 3D reconstruction.

# Why is it an issue if it's working ?

- 1) *Not very elegant !* said the computer engineer
- 2) *Nest of bugs !* said users
- 3) *Hard to maintain !* said the programmer
- 4) *Very resources-intensive to change !* Said our boss
  - Ex: HDF5 (or whatever file format) migration
    - implementation in SPEC
      - could be done by SPEC maker
    - migration of all our macros :
      - 20 000 files to inspect ?
      - 1 year.man just to check which macros contain saving related code...



# Why is it an issue ?

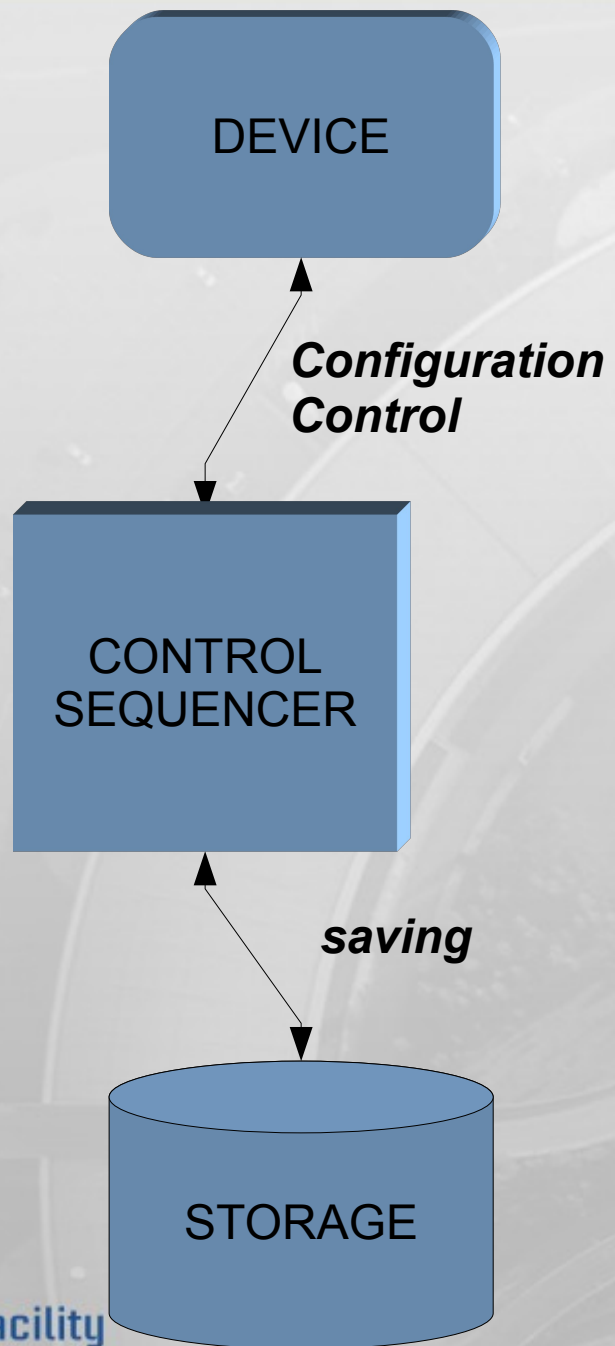
5) *Difficult to combine control and on-line tools !* said all

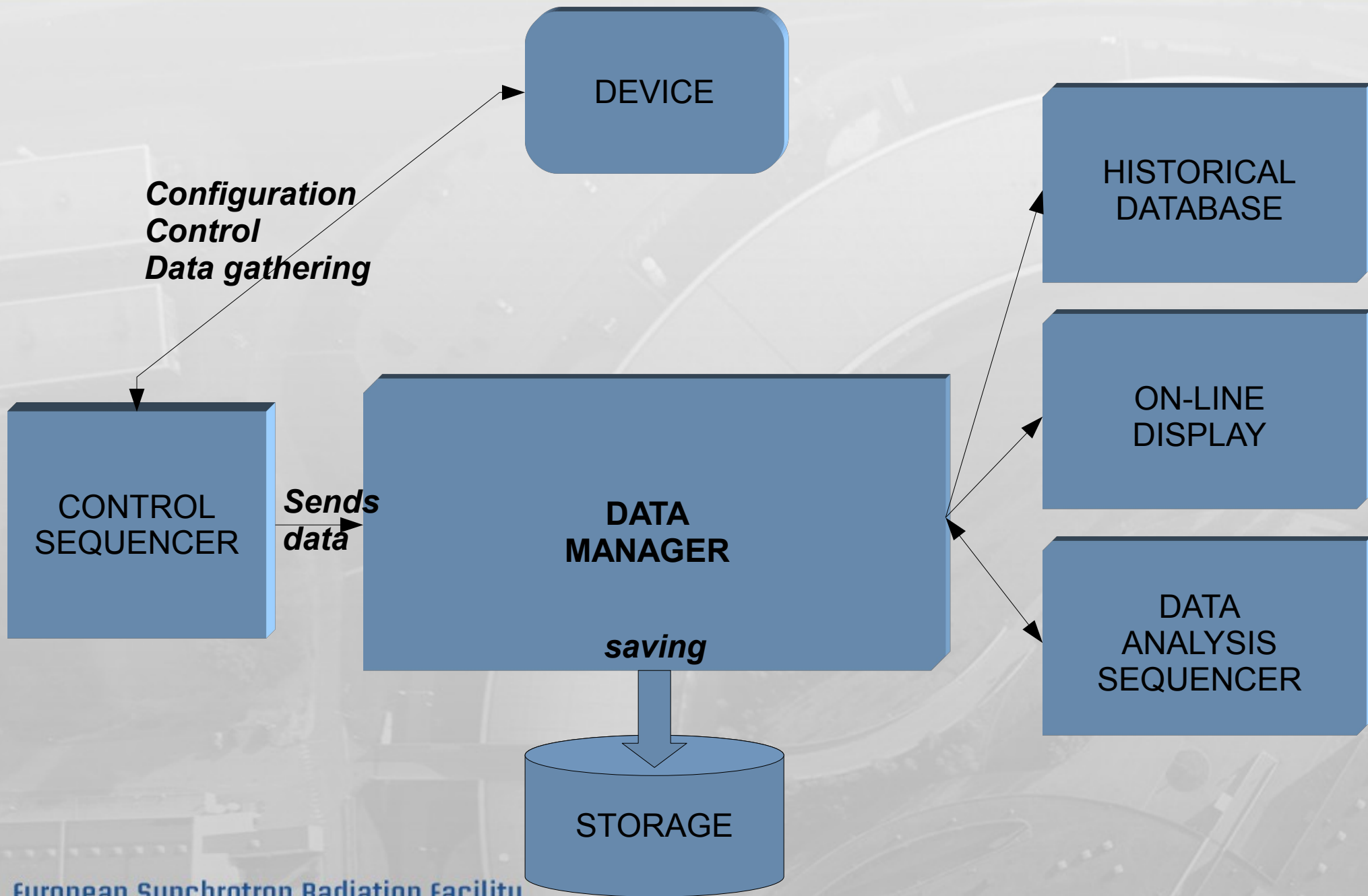
- Data analysis tools do not know where to find data and how to safely access it
  - Ex : Read data in memory of a device server or on disk ?
  - Ex : Is the image really fully written on disk ?
    - Wait → latency of several seconds
- We must ensure that analysis tools have access to all needed data
  - Ex : do not forget to save *ring current* even if it is useless at experiment time.

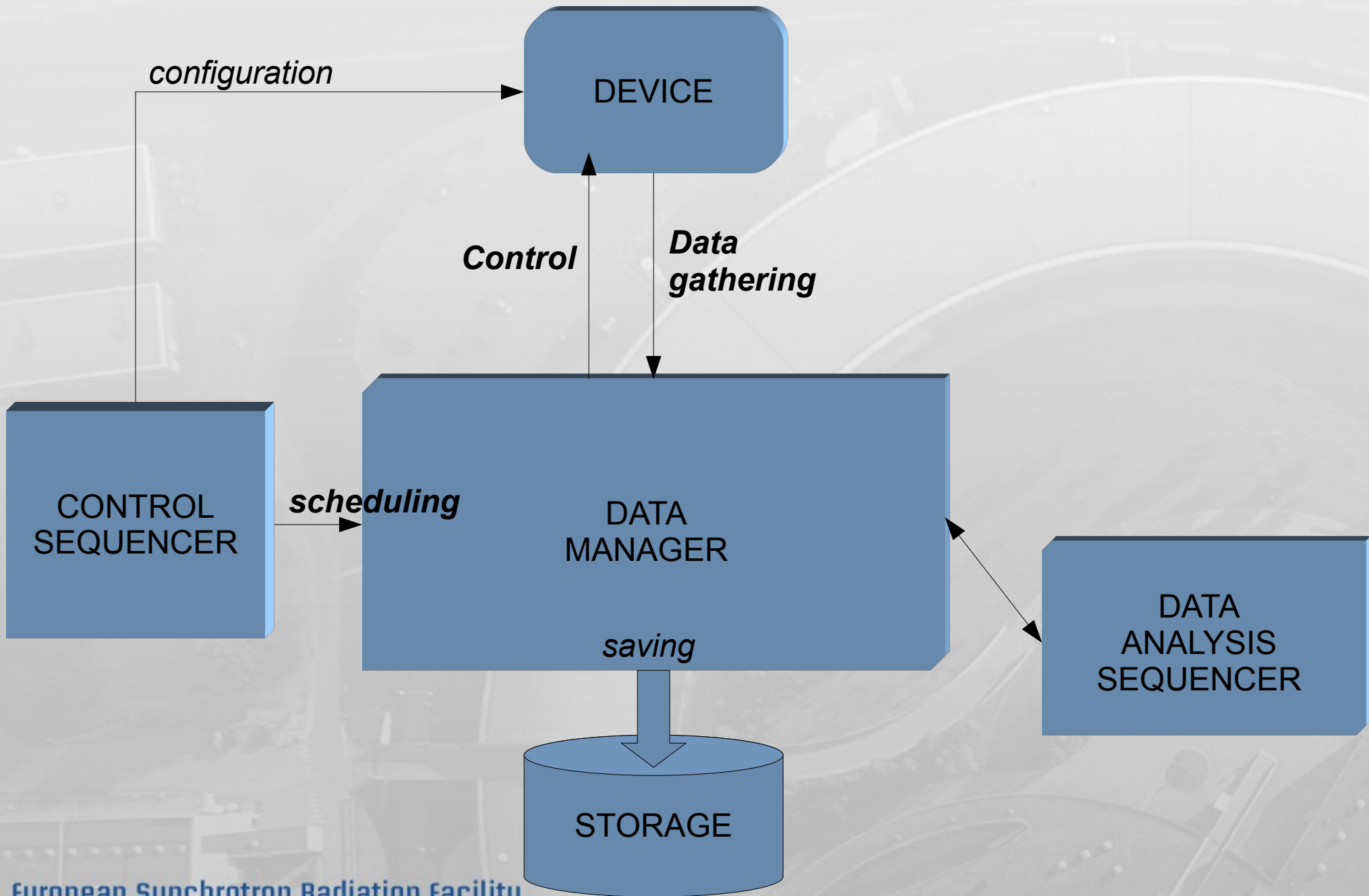
# Why is it an issue ?

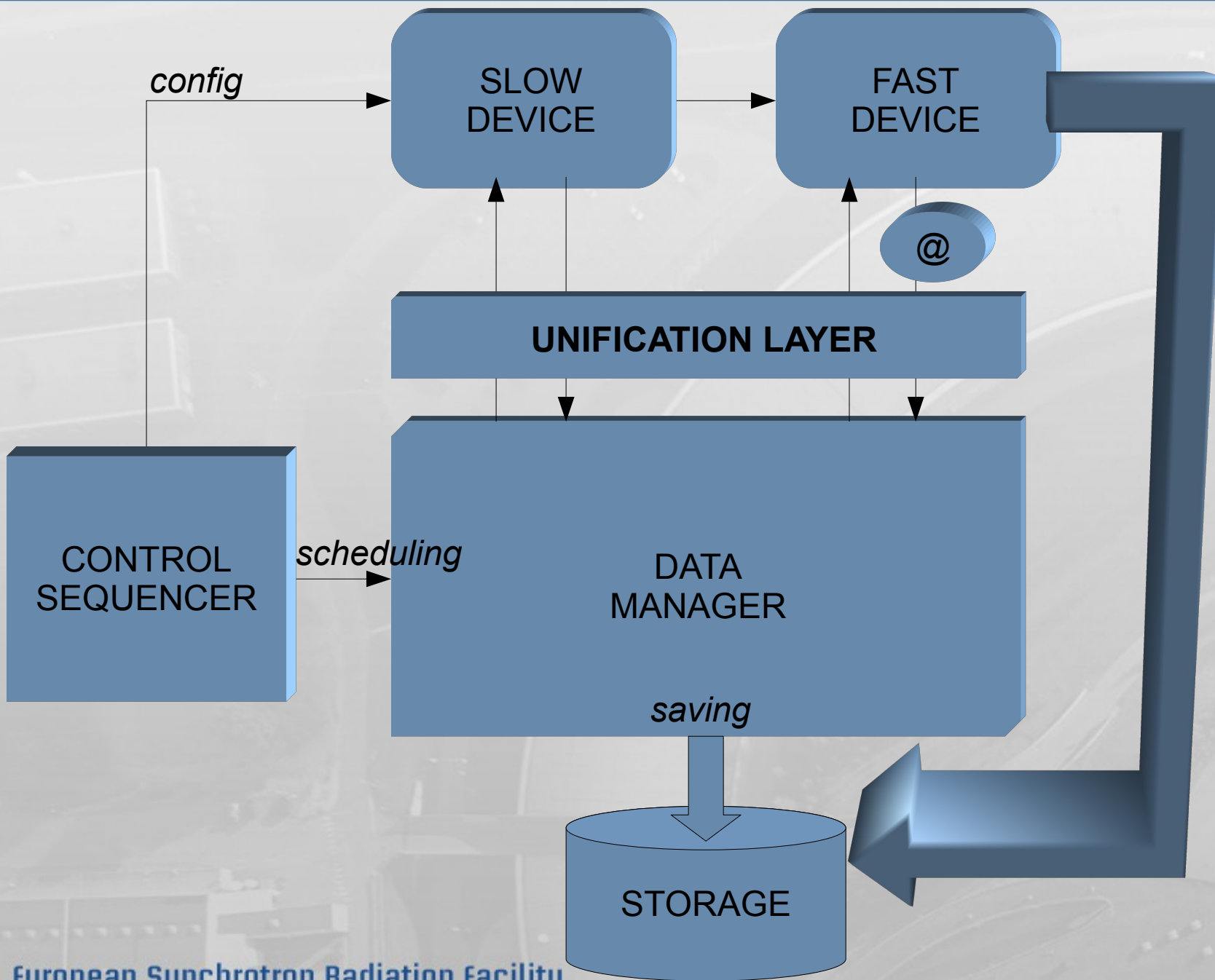
6) *Tools do not have knowledge of the semantic of data !* said the analysis tool maker

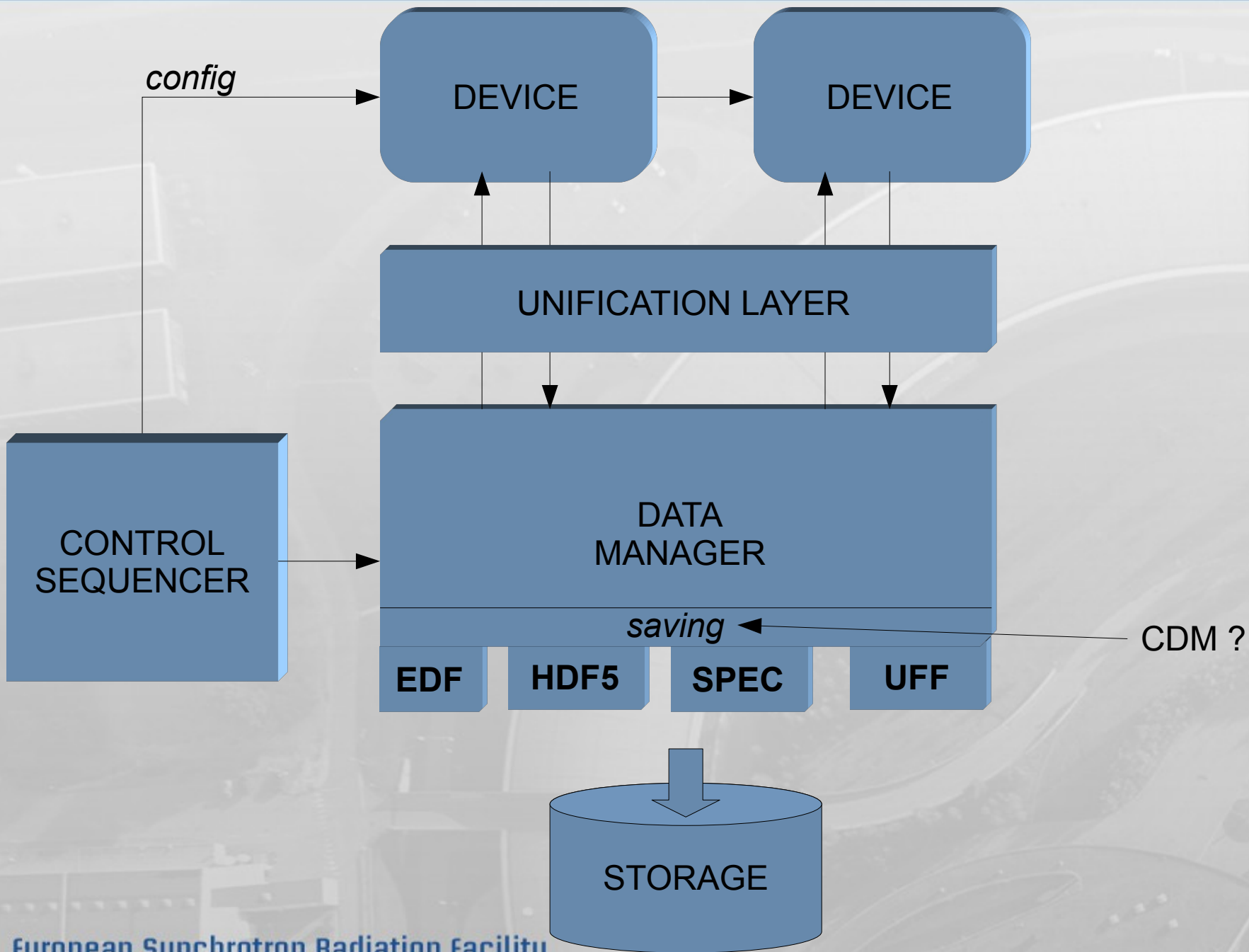
- A raw data is typically a (short|long|float) (0|1|2|3)d array.
- The signification of the raw data must be known by the analysis tool to be treated with least user intervention.
- This signification depends on the experiment
- A sequence can produce data with different semantics
  - a same scan can be used for a tomography or for mapping the analysis part will be different.





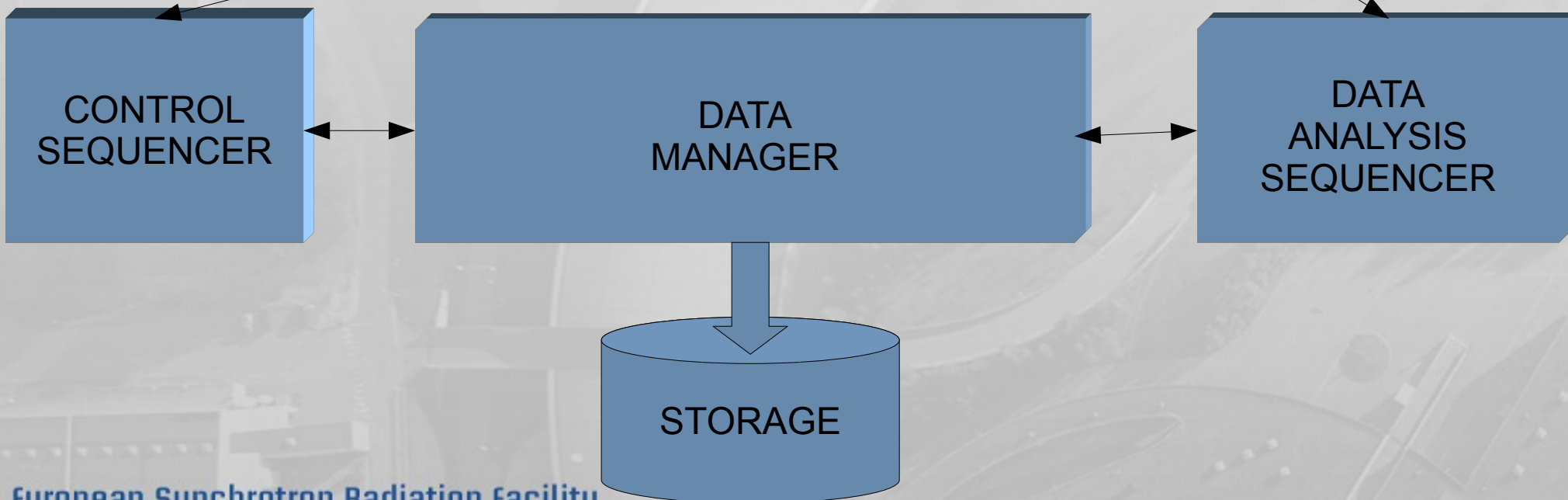






```
sequence : "EFF"  
  → paramSet : "Experiment Description"  
    -user : "toto"  
    -exp name : "tutu"  
  → acquisition : "Dark for white"  
  → acquisition : "Dark for data"  
  → scan : "Energy Scan"  
    → paramSet : "Scan parameters"  
      -energy start = 3.5  
      -energy end = 4.5  
      -steps = 100  
    → N sequences :  
      → acquisition : "MonoPos"  
      → acquisition : "White before"  
      → acquisition : "Data"  
      → acquisition : "White after"
```

DATA STRUCTURE  
DESCRIPTION  
(experiment dependent)





DATA  
STRUCTURE  
CREATION  
TOOL



```

sequence : "EFF"
├── paramSet : "Experiment Description"
│   ├── -user : "toto"
│   └── -exp name : "tutu"
├── acquisition : "Dark for white"
├── acquisition : "Dark for data"
└── scan : "Energy Scan"
    ├── paramSet : "Scan parameters"
    │   ├── -energy start = 3.5
    │   ├── -energy end = 4.5
    │   └── -steps = 100
    └── N sequences :
        ├── acquisition : "MonoPos"
        ├── acquisition : "White before"
        ├── acquisition : "Data"
        └── acquisition : "White after"
    
```



ANALYSIS  
WORKFLOW  
DESCRIPTION  
TOOL

# Data Manager prototype

- Briefly tested in December on ID21 for a full field EXAFS experiment in close collaboration with data analysis unit.
  - Dedicated PyTango Device server
  - SPEC as a sequencer
  - dialog device server to send data to be treated to EDNA
    - Normalization
    - Image Alignment
- Simplicity in a key point

# Conclusion and Future work

- To finish the prototype
- To continue running prototype tests
- To check pertinence of this approach on various experiments
- Did you said performances ?
- Continue gathering good ideas