InCAEM

Data Infrastructure

Sergio Vicente Molina

Finançat per:

Unió Europea
Fons Europeu
Next Generation

GOBIERNO DE ESPAÑA

Plan de Recuperación, Transformación y Resiliencia

Next Generation Catalunya

Generalitat de Catalunya
Departament de Recerca i Universitats

WP4: Infrastructure & methods for data analysis

G. Merino (PIC), P. Ordejón (ICN2), A. García (ICMAB), M. Eriksen (PIC), F. Torradeflot (PIC), V. Acín (PIC)

Nowadays the computational capacity is a key success factor in applied science

WP4&5 teams work together to provide the required IT infrastructure and data management platform

WP5: Infrastructure & methods for in situ data analysis

S. Vicente (ALBA), A. García (ICMAB), J. Otón (ALBA), N. Soler (ALBA), G. Rosas (ALBA)

Finançat per:

Unió Europea
Fons Europeu
Next Generation

GOBIERNO DE ESPAÑA

Plan de Recuperación, Transformación y Resiliencia

Next Generation Catalunya

Generalitat de Catalunya
Departament de Recerca i Universitats

Physical location

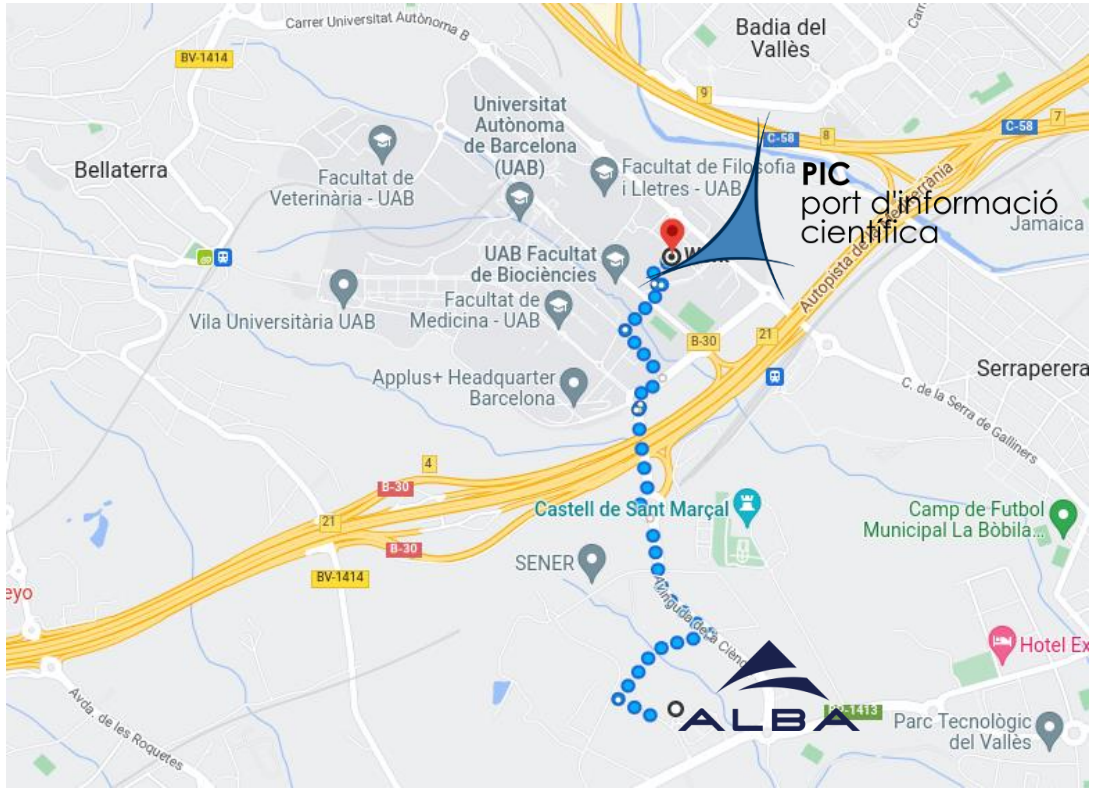
- The ALBA synchrotron is the location where the microscopes will be installed.

- PIC is a datacenter for the LHC experiment at CERN, the magic telescopes, the Euclid satellite and several other projects.

- Physically close, about 30 minute walk.

- Excellent network connection between sites (dark fiber).

- Both institutions already collaborate in other IT infrastructure initiatives (i.e. online remote backup)



Finançat per:

Unió Europea
Fons Europeu
Next Generation

GOBIERNO DE ESPAÑA

Plan de Recuperación, Transformación y Resiliencia

Next Generation Catalunya

Generalitat de Catalunya
Departament de Recerca i Universitats

Overall layout

STEM and SPM microscopes experiments

Data acquisition

Correlative beamline experiments

In-situ Data Analysis platform

Network connection via CSUC - Rediris

Offline Data Analysis and Simulation platform

Data annotation, storage, archive and long term preservation

Data annotation, storage, archive and long term preservation

Finançat per:

Unió Europea
Fons Europeu
Next Generation

GOBIERNO DE ESPAÑA

Plan de Recuperación, Transformación y Resiliencia

Next Generation Catalunya

Generalitat de Catalunya
Departament de Recerca i Universitats

Data management

• **In-situ data analysis**. First quick analysis to provide user feedback and enable optimization of sample condition. Run mainly at ALBA

• **Offline data analysis.** Rerunning pipelines, analyze data and train/deploy deep learning models. Run mainly at PIC

• **Theoretical computation and simulations.** It will share the resources of the data analysis platform, optimizing the use of the computational resources taking profit of the data analysis idle time

• **Long term storage.** Preserve data for the duration of the project and beyond. Using tape media in two different locations.

•**FAIR data.** Capture appropriate metadata, use of a data catalogue and enable remote open access to that data, with proper DOIs.

Finançat per:
Unió Europea
Fons Europeu
Next Generation
GOBIERNO DE ESPAÑA
Plan de Recuperación, Transformación y Resiliencia
Next Generation Catalunya
Generalitat de Catalunya
Departament de Recerca i Universitats

Current status

- **Survey and interviews to scientists to gather requirements almost finished**

- **Initial conclusions:**
  - **STEM :** Identified possible cutting-edge experiments that could consume huge amounts of storage, network and computing capacity. The volume of data would represent tenths of TB of raw data, but in exponential growth after analysis execution. In those cases, specific compression algorithms will be required
  - **AFM / STM :** Instruments and experiments are not so computing intensive as STEM. They should be managed with minor upgrades of the existing IT platforms
  - **Theoretical computation and simulations :** Require a relevant CPU, RAM and GPU capacity but not storage
  - **Correlative beamlines experiments :** Could require minor upgrades of the existing IT platforms for beamlines but important developments of software and tailoring of data management pipelines

- **Initiating data management plans (DPMs)**

- **Initiating the detailed definition of the IT infrastructure** to identify the required **call for tenders**

Finançat per:
Unió Europea
Fons Europeu
Next Generation
GOBIERNO DE ESPAÑA
Plan de Recuperación, Transformación y Resiliencia
Next Generation Catalunya
Generalitat de Catalunya
Departament de Recerca i Universitats

Budget

WP4: Infrastructure & methods for data analysis 1.100 K€

WP5: Infrastructure & methods for in situ data analysis 1.100 K€

Example of platforms to be upgraded:

- Servers with CPU and GPU for increasing the HPC capacity
- Data network switches for high speed communications
- Data storage cabins for online storage
- Tape libraries for long term preservation
- Firewalls for network security
- Workstations for in-situ quick data analysis
- Software licenses (for example, for efficient download of huge amounts of data)
- …

Finançat per:
Unió Europea Fons Europeu Next Generation
GOBIERNO DE ESPAÑA
Plan de Recuperación, Transformación y Resiliencia
Next Generation Catalunya
Generalitat de Catalunya Departament de Recerca i Universitats

Conclusions

- Data analysis and management is done in **collaboration between ALBA and PIC**

- The In-CAEM compute infrastructure takes profit of the **current infrastructure** and previous experience, but will require **important capacity upgrades**

- Computing infrastructure optimized for data processing that will support both **experimental analysis and simulations**

- **Deep learning (AI) workloads** is being integrated from the start for e.g. feature detection in electron microscopy imaging

- **Data Management Plans (DPMs)** are essential for the correct management of the data from the initial stage, the data acquisition, to the final stage, the long term preservation and publication using the FAIR principles

Finançat per:
Unió Europea
Fons Europeu
Next Generation
GOBIERNO DE ESPAÑA
Plan de Recuperación, Transformación y Resiliencia
Next Generation Catalunya
Generalitat de Catalunya
Departament de Recerca i Universitats