SCTB review:
Computing and data analysis

Nicolas Soler - 09/11/2021
Computing Division

The Computing division

Computing & Controls

Oscar Matilla
Computing Division Head

Concepción Girbau
Secretary

New
Feb 2021

Controls and DAQ

Guifré Cuní
Controls Section Head

Electronics

José Ávila
Electronics Section Head

IT Systems

Toni Pérez
IT Systems Section Head

MIS

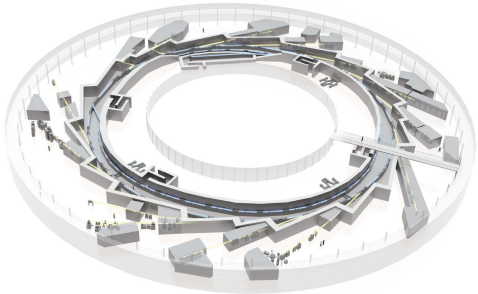
MIS Section Head

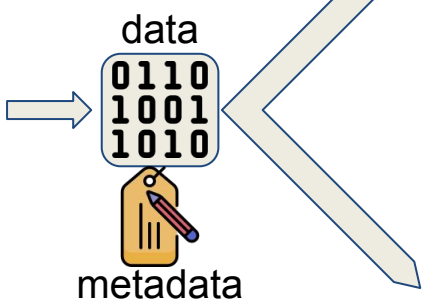
SDM

Nicolas Soler
SDM Section Head

2021: Scientific Data Management

A new section inside the Computing & Controls division

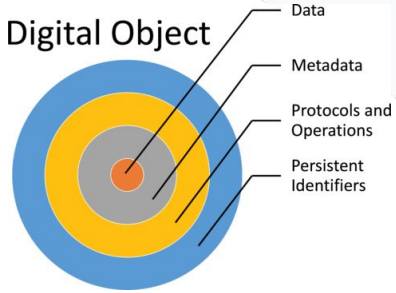




Data processing

Data reusability

- Metadata ingestion
- Provenance
- Persistent identifiers
- Catalogue
- DAaaS



4 people hired with hybrid science / computing profiles (+1 more in the beginning of 2022)

- Beamlines support
- Occasional accelerators support

https://www.nist.gov/programs-projects/facilitating-adoption-fair-digital-object-framework-material-science

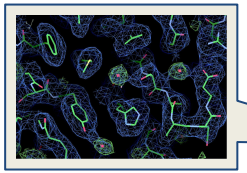
Data availability

The data supporting this study can be made available from the corresponding author upon request.

Towards FAIR synchrotron data

panosc

ExPaNDS



User: connection from local workstation or home laboratory

re-process

Data analysis as a service (WP4)

- EOSC Service portal
- PaN portal
- Containers
  Virtual machines
  Remote desktops
  Jupyter notebooks
- HPC
- Reference datasets

Use
re-use

Process
raw data

Annotate &
store processed data + provenance

FAIR-ready data (WP2)

- **Certified** (meta)data repository
- Standard metadata framework
- FAIR Data Management Plan (**DMP**) for each technique
- Persistent Identifier (**PID**) for data, instrument, software, sample)
- Common **data policy** framework

Data catalogues (WP3)

- **API / UI**
- **EOSC hub**
- **Metadata ontologies** (shared vocabulary)
- PaNOSC **federated catalogue**
- **Metadata formats** (NeXus)
- **Catalogues** (eg iCAT, SciCat)

Data reduction / compression

- Metrics
- New algorithms
- Software vs hardware
- Technique-specific
- Lossy vs non lossy
- Meta-compressors (ML)

LEAPS INNOVATION (WP7)

Implementation of a data catalogue at ALBA

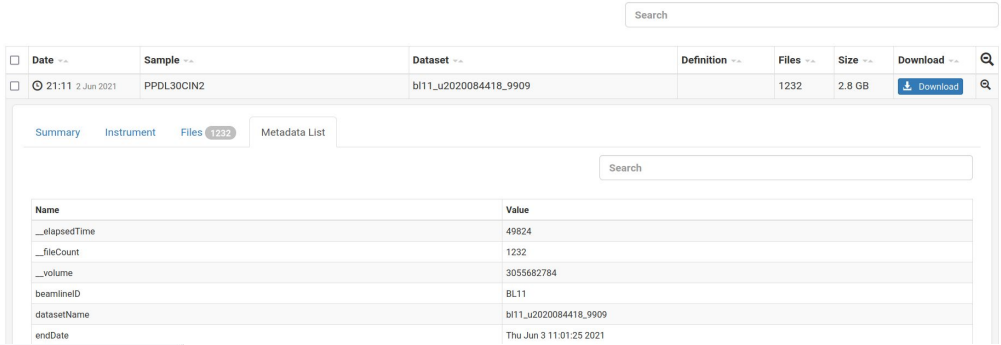
EOSC portal

PaNOSC Federated Metadata catalogue

- iCAT
- iCAT
- SciCAT
- …..

Experimental data, metadata, derived data and related info (i.e PIDs)

European facilities data catalogues



B2FIND

doi .org

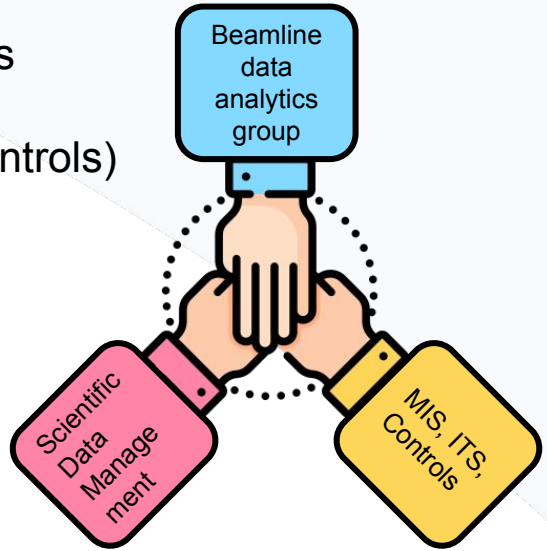
OpenAIRE

- Data have an embargo period of 3 years
- iCAT already in use for NCD-SWEET
- Next: Implementation in MIRAS and MSPD
- Open externally by the end of the year

2021: Scientific Data Management

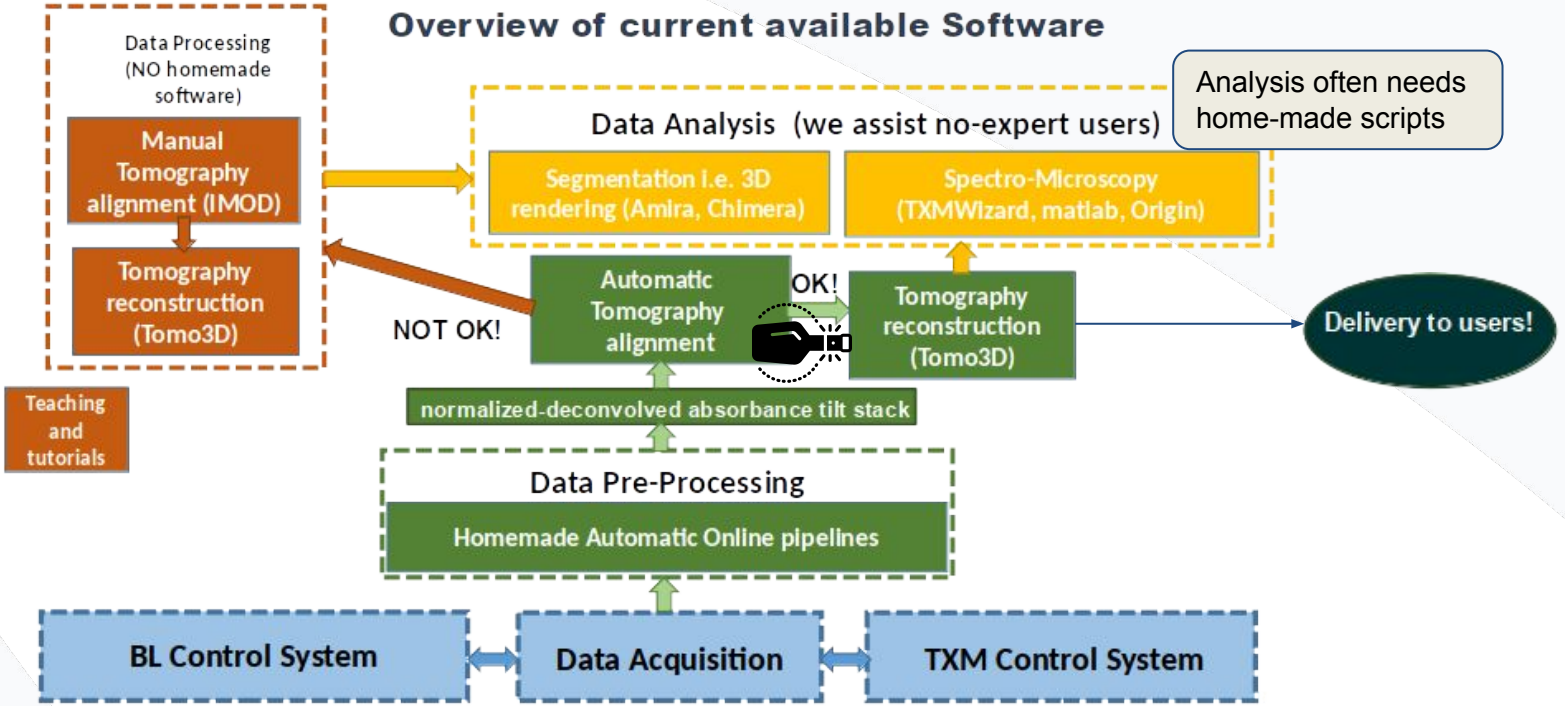
BUILDING A SYNERGIC SOFTWARE DEVELOPMENT PROCESS BETWEEN COMPUTING ENGINEERS AND METHOD DEVELOPMENT SCIENTISTS AT THE BEAMLINE

- Building downstream data processing and analysis pipelines
- Real-time data processing triggering and feedback (with Controls)
- Data visualization tools and GUI
- Databases and APIs
- Data reusability (ExPaNDS, with MIS and Controls)
- **Collaborating in data analysis methods development**



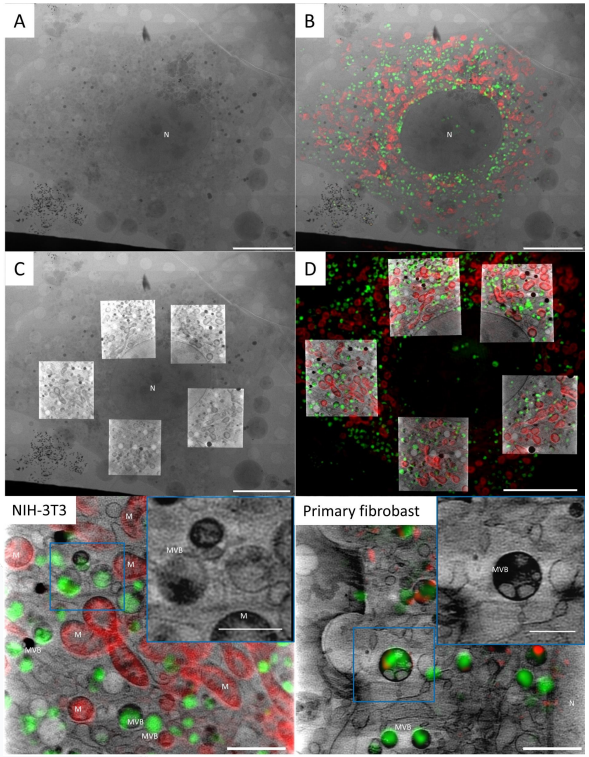
MISTRAL data flow (cryo-SXT)

(from A.Sorrentino' slide



Example of data analysis

Effect of an antifibrotic novel therapeutic agent at the cellular level



Chemical Science

ROYAL SOCIETY OF CHEMISTRY

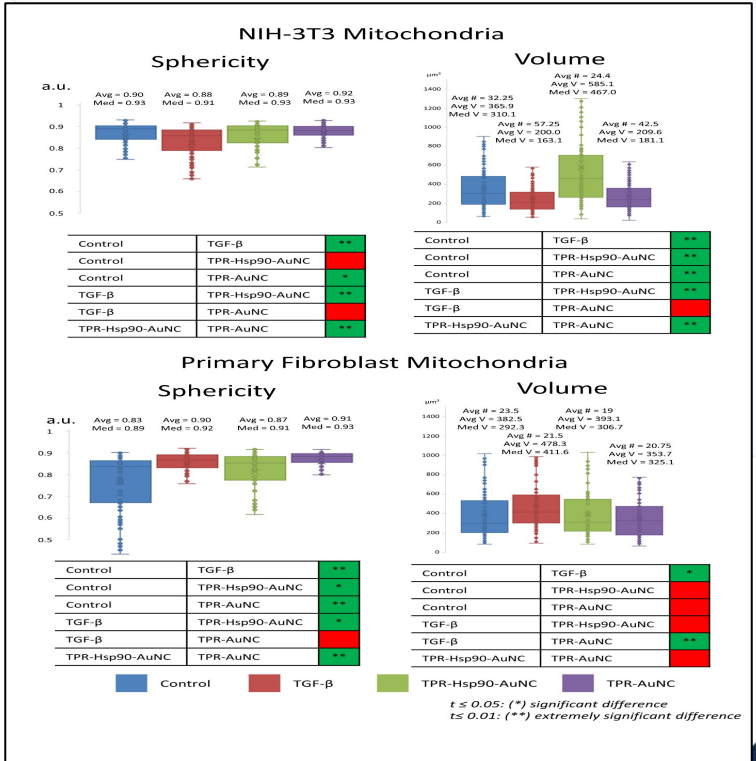
EDGE ARTICLE

Cite this: DOI: 10.1039/d1sc04183e

All publication charges for this article have been paid for by the Royal Society of Chemistry

Correlative 3D cryo X-ray imaging reveals intracellular location and effect of designed antifibrotic protein–nanomaterial hybrids†

J. Groen, [a] A. Palanca, [bc] A. Aires, [d] J. J. Conesa, [ae] D. Maestro, [b] S. Rehbein, [f] M. Harkiolaki, [g] A. Villar, [bh] A. L. Cortajarena [*di] and E. Pereiro [*a]



MISTRAL DP/DA: GAP analysis

DP feature	Current setup	Problem / disadvantage	Mitigation / future setup
User interface	user interface	Scripts-based	GUI / workflow manager
Preprocessing	Custom pipelines (single/multiple images per angle for tomos, spectroscopy 2D & 3D)	has to be launched manually after data acquisition	Automatic triggering to be installed
Alignment (tomo & spectra)	IMOD, EFTEM-TOMOJ & 2D correlation Image J or python (2D spectroscopy)	presence of fiducials required but hardly controllable	Optical flow, TomoPy
Reconstruction	SIRT (Tomo3D), ART of TOMOJ	ok but could be GUI-monitored	followed on workflow manager
Segmentation & analysis	AMIRA, Chimera, SuRVoS	Not enough automation?	ML approaches
Correlative microscopy 3d	in progress	to test	*Icy- eC-CLEM?*
Data access, reprocessing	via sFTP, HDD	preservation, provenance	Data catalogue, DAaaS

MISTRAL data processing & analysis

- Bring more **automation** in the current data processing pipeline. Bring graphical interfaces and data workflows where needed. Improve current pipelines.

- Install a robust, parallelized xy **fiducialess alignement** correction program (e.g Tomopy) that could be use by default (whether or not fiducials are present).

- Provide software support for **correlative microscopy** (cryoET and cryo3DSIM volume superposition) and **segmentation methods**

- Together with on-site experts, assist in the development of **analysis** software.

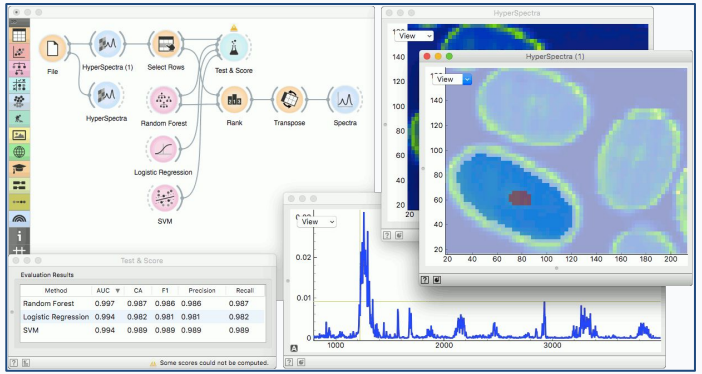
FAXTOR hard X-ray µCT: *Future issues to consider*

- Workflow manager and previews (GUI)
- preprocessing: from raw data to phase retrieval, (normalisation etc).
- computer reconstruction (aiming at real-time, parallelized on the cluster)
- image registration (re-aligning / shifting)
  - on angular projection
  - on reconstructed volume
- Machine learning for low dose, segmentation, denoising, missing angles etc
  - on angular projection
  - on reconstructed volume
- Managing and visualizing big data volumes (size: few GB to several TB)
  - distributed memory visualization (NVIDIA index, paraview plugin)
  - Data selection (automated?), Database of scans (iCAT)
  - Data compression (probably lossy)

Cell and tissue biology beamlines:

How can SDM help?

MIRAS IR spectroscopy & microscopy



- Move from proprietary to **open source solutions** (e.g Orange, Quasar). Join the IR quasar network (SOLEIL, Elettra, Sirius, Canadian light source etc), especially bringing **data workflows**, **machine learning** capabilities and nice feature to explore **imaging**/hyperspectral data.
- Improve the data analysis as a service by switching to these solution.
- Provide advice for **statistical methods** (eg t-SNE instead of PCA)
- **Multimodal** experiments: correlation with other types of spectroscopies

quasar

biolab

SOLEIL SYNCHROTRON

Elettra Sincrotrone Trieste

Canadian Light Source | Centre canadien de rayonnement synchrotron

BioSpec Norway

LNLS Brazilian Synchrotron Light Laboratory

PHOTOTHERMAL SPECTROSCOPY CORP

orange

Summary

- SDM has been created 9 months ago to support Experiments with data processing, analysis and visualization needs as well as with future data reusability (FAIRness)
- The first 2 engineers started in September, 1 in November, 1 more to come until in Jan. 2022, 1 to be hired next year.
- SDM will coordinates the metadata ingestion scheme and catalogue, as well as the transition towards data analysis as a service for remote users.
- Computing engineers with scientific background will work hand in hand with data analytics beamline method developers to:
    - produce and maintain data processing & analysis pipelines, bringing expertise in software development and other relevant technologies (i.e machine learning.
    - Provide support for integration of heterogeneous data.