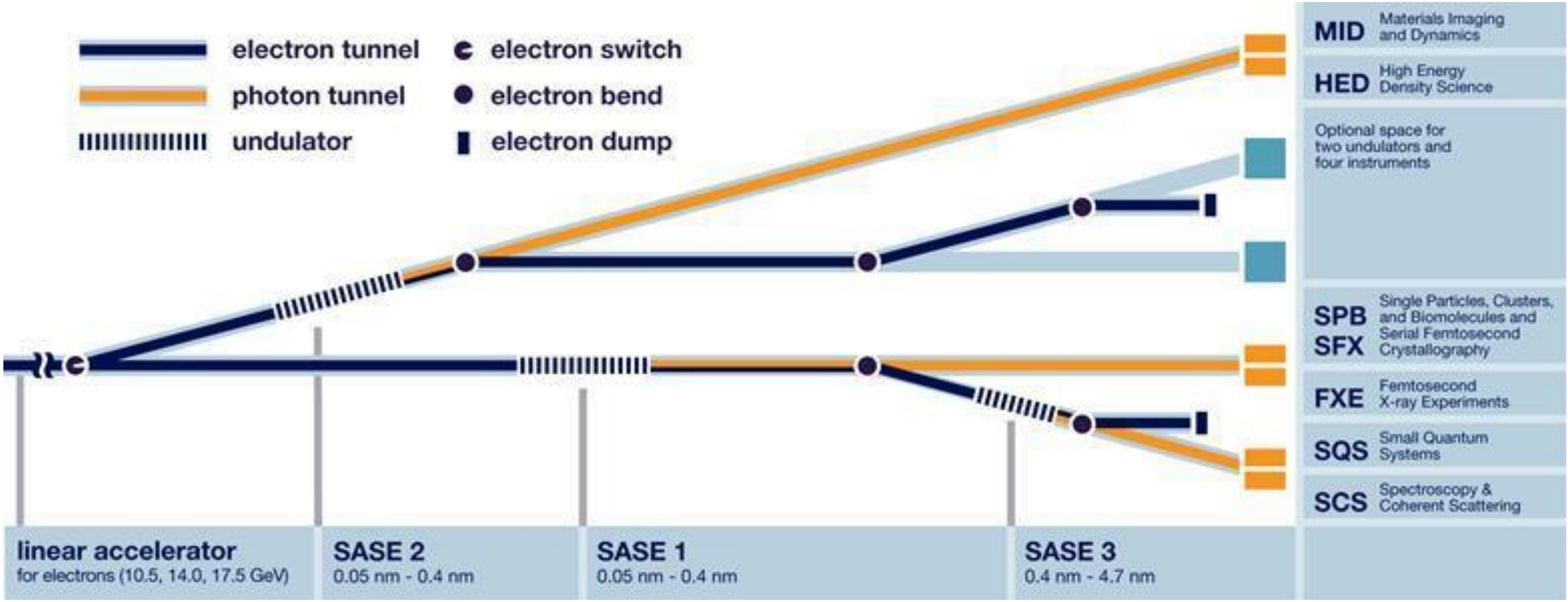
Jupyter for Reproducible Science at Photon and Neutron Facilities

Authors: Robert Rosca, Hans Fangohr
European X-ray Free Electron Laser GmbH
PaNOSC - Photon and Neutron Open Science Cloud

Reading, 15th October 2019

robert.rosca@xfel.eu

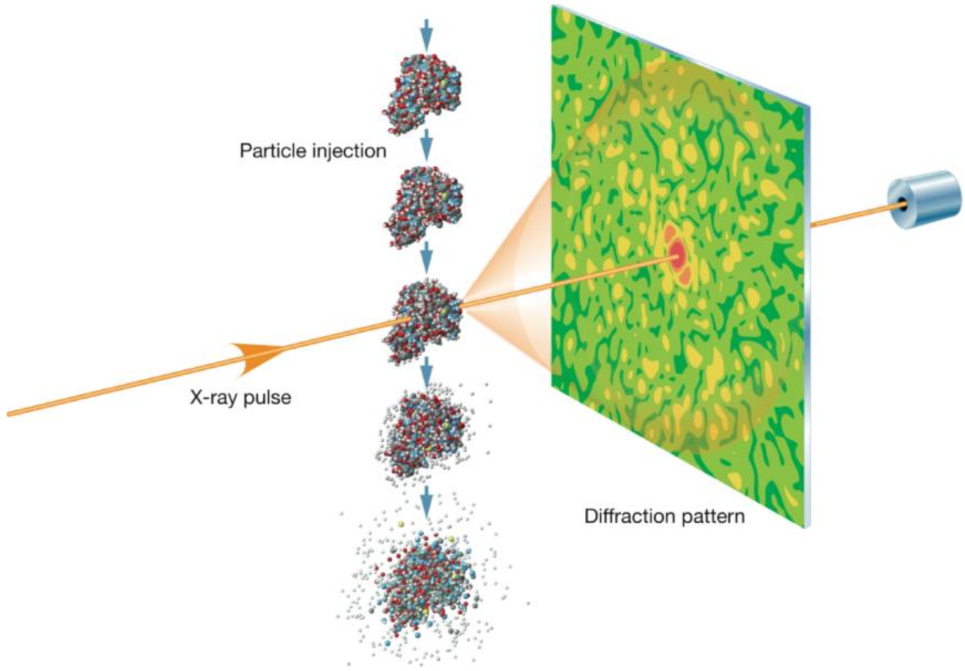
Intro to EuXFEL



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823852.

Intro to EuXFEL

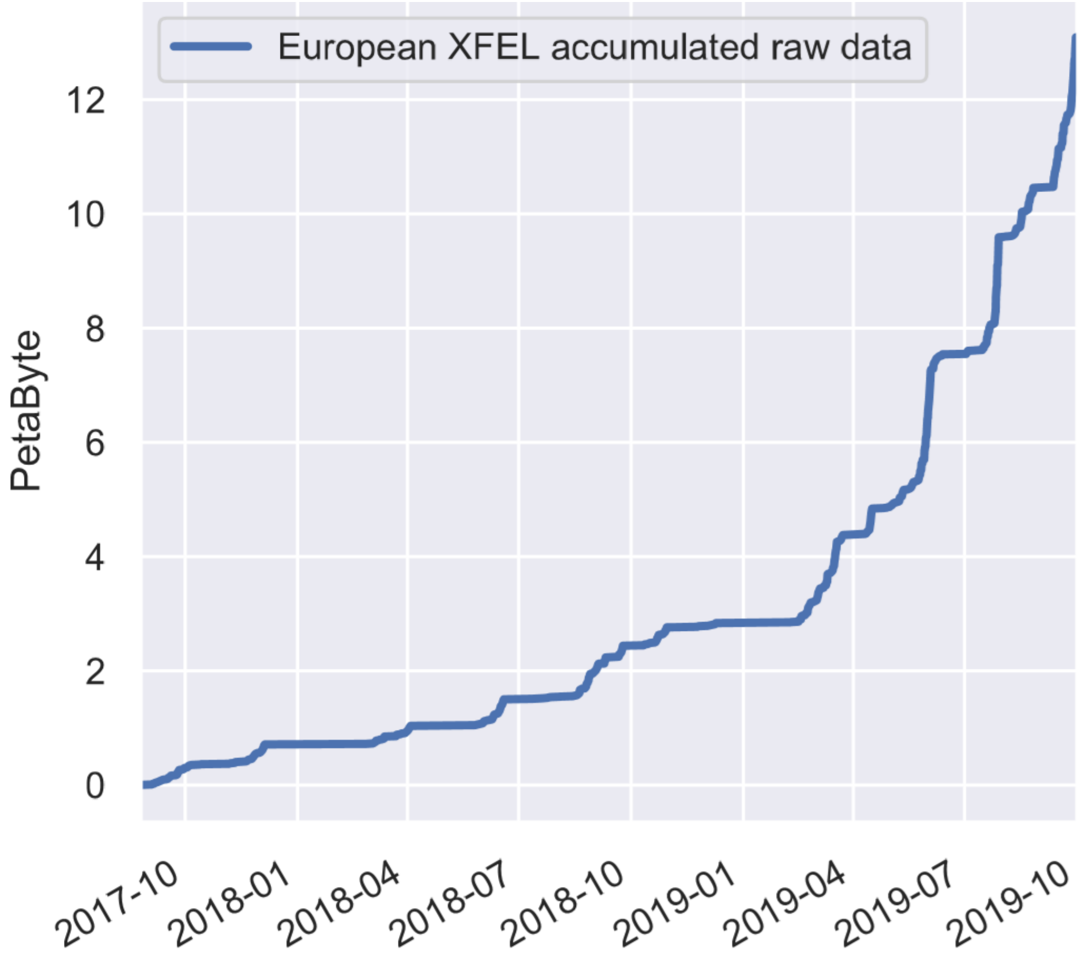
- 27 000 X-ray pulses per second (in 10 Hz bursts)
- European XFEL: use short-wavelength photons to image small things
  - Data volume
    - Typical detector: 1 million pixels, each using 2 bytes
    - up to 27 000 X-ray pulses per second
    - → 2 byte * 1 000 000 * 27 000 / s = 54 GB/s
    - → 194 TB/h (theoretical peak)



Intro to EuXFEL

- Data collected by experiments is massive
- Up to hundreds of TiB per run, and dozens to hundreds of runs per experiment
- Users cannot bring this data back to their facilities

- Other facilities have similar problems



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823852.

Other Facilities

Data / yr	ILL	ESRF	CERIC	XFEL	ELI	ESS
2019	200 TB	8 PB	1 PB	3 PB	< 1 PB	0
2023	600 TB	50 PB	15 PB	100 PB	10 PB	< 1 PB

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823852.

Photon and Neutron Open Science Cloud – team effort

- Sandor Brockhauser (EuXFEL)
- Aidan Campbell (ESRF)
- Hans Fangohr (EuXFEL)
- Andy Götz (ESRF)
- Jamie Hall (ILL)
- Jerome Kieffer (ESRF)
- Thomas Kluyver (EuXFEL)
- Eric Pellegrini (ILL)
- Jean-François Perrin (ILL)
- Carlos Reis (CERIC-ERIC)
- Thomas Rod (ESS)
- Robert Rosca (EuXFEL)
- Jesper Selknaes (ESS)
- Krzysztof Wrona (EuXFEL)

panosc
photon and neutron open science cloud

- ESRF: European Synchrotron Radiation Facility
- ILL: Institut Laue-Langevin
- EuXFEL: European X-ray Free Electron Laser Facility
- ESS: The European Spallation Source
- CERIC-ERIC: The Central European Research Infrastructure Consortium
- ELI: Extreme Light Infrastructure
- EGI: European Grid Infrastructure

Data Analysis for Open Science

- FAIR (Findable, Accessible, Interoperable, Reusable) data is central for Open Science
- Data analysis extracts the meaning from the data
- Publications based on data
  - Data sources should be known
  - Central findings (figures, tables, numbers) should be reproducible

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823852.

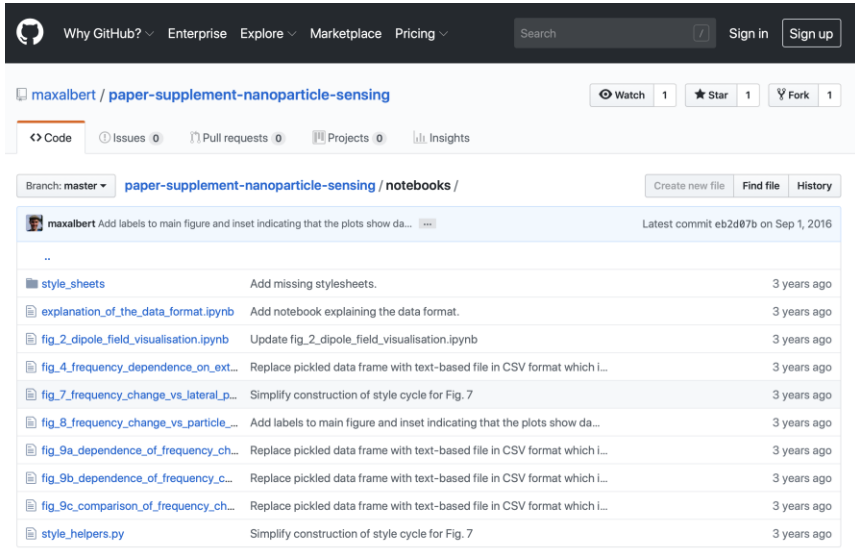
Jupyter Notebook for Open Science

- Combination of code, output and annotation *in one document*

- If used appropriately, makes publications *reproducible*
  - For example: one notebook per figure in publication (examples: [1], [2])

- Notebooks from reproducible publications make the work *re-usable*
  - Currently, lots of time is used by researchers to repeat the work of others, before they can advance science.

[1] Appl. Phys. Lett. 109, 122401 (2016), https://doi.org/10.1063/1.4962726,
https://github.com/fangohr/paper-supplement-2016-dmi-nanocylinder-hysteresis
[2] Nanotechnology 27, 455502 (2016), https://doi.org/10.1088/0957-4484/27/45/455502
https://github.com/maxalbert/paper-supplement-nanoparticle-sensing



Solution architecture for PaNOSC vision:

- Finding data:
  - Web interface with data base of experiment metadata
- Exploring and analysing data remotely (in ‘the cloud’):
  - JupyterHub serving relevant notebooks
  - Move data analysis code into the notebook
  - Remote desktop in browser, connected to Desktop of virtual machine

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823852.

PaNOSC Use Case 1: reproducibility and re-usability published results

- For a given publication based on facility data, users can
  - Find the data (through the EOSC web portal or URL/DOI in paper)
  - Access the data through web portal
  - Inspect the data analysis (notebook) that led to key figures / statements in the publication
  - Re-execute the data analysis through (→ reproducibility)
  - Modify and extend the notebook (→ reusability)

- Users may include scientists, interested public, journal editors and reviewers, representatitves from research councils, . . .

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823852.

PaNOSC Use Case 2: enable new data analysis on existing data sets

- Users can
  - Search and find data sets from experiments through web portal
  - Access the data through web portal
  - Choose from appropriate selection of data analysis tools (=Jupyter Notebook templates)
  - Execute the notebook
  - Modify and extend the notebook

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823852.

Challenges 1: interoperability between facilities

- Different facilities
  - Currently 6 facilities involved
  - Generally use different ways to store metadata
  - Common way of classifying data sets (and experiment types)?

- Data scale
  - For some data sets, the data cannot be moved to the compute resource

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823852.

Challenges 2: data analysis in Jupyter Notebooks

- Analysis in Notebook
  - Computational environment – software (containers, singularity?)
    - Need to provide the right computational environment for each analysis type
    - How can we maintain computational environments in the future (Binder-like?)
    - Extending up to the life-time of publications and data sets
  - Which analysis is appropriate for data set → classification of data sets
  - Making analysis capabilities available in the Jupyter Notebook
    - Command line driven and Python based computation
    - GUI-based tool more difficult / impossible
  - What to do with resulting analysis?
  - Some analysis notebooks require significant HPC resources (execute jobs from notebook?)
  - Computational environment – hardware, GPUs?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823852.

Challenges 3: policy, organisation, and culture

- Concurrent development with EOSC hub
- Complicated access rights for data sets
  - Data policies with embargo period
- Publication use case:
  - Requires collaboration with scientists
    - Preparation of data analysis notebooks
  - Social / cultural challenge
  - Helped by changing metrics and expectations from funding bodies and journals
  - Research facilities can lead by example

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823852.

Work Done so Far

- Things are still quite uncertain
- Time spent on researching tools and viability, currently investigating:
  - Singularity for containers
  - MyBinder-like setup for remote users
  - Jupyter Kernels running out of Singularity for local users
- Website mockup created
- Work done on remote desktop services for GUI-heavy applications

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823852.

PaNOSC

search

Filters

Data Type

Simulation	521
Experiment	2560
Derived	423

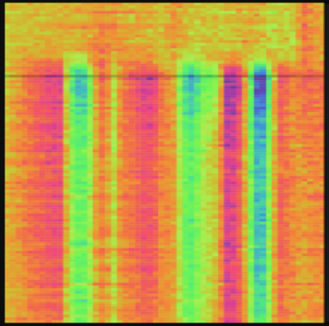
Field

X-Ray Sources	368
Plasma Physics	49
Ion Acceleration	76
Electron Acceleration	85
Material and Biomolecular Applications	122

Technique

X-ray phase contrast imaging	59
X-ray Diffraction	45
X-ray absorption spectroscopy	85
Coherent Diffractive Imaging	26
Atomic, Molecular and Optical Science	736
Soft X-ray Materials Science	48
Pulsed Radiolysis	29
WW pump-probe	47
X-ray Phase contrast imaging	14
X-ray fluorescence	238
Absorption spectroscopy, WDM@10Hz	45

Datasets



Time-resolvent spectroscopy - run 1-52

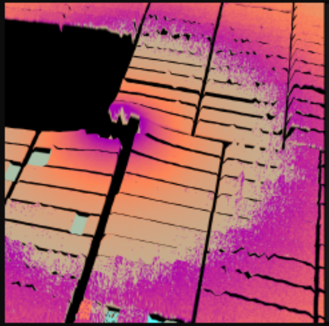
RP4-SRS focuses on time-resolvent spectroscopy experiments in the full range of frequencies from IR to UV. Users can measure samples as varied as solid state crystals, or proteins in their natural environment. Time-resolved spectroscopy is the collection of techniques that are used to examined the dynamic processes of materials and chemicals upon illumination with a pulsed laser...

Dataset | X-ray Spectroscopy | Pulsed Radiolysis | All Tags 8

Created 2019/03/15
Size 328 MB
Views 3

jupyterlab

launch VM



Two-color XUV+NIR femtosecond photoionization of neon in the near-threshold region

RP4-SRS focuses on time-resolvent spectroscopy experiments in the full range of frequencies from IR to UV. Users can measure samples as varied as solid state crystals, or proteins in their natural environment. Time-resolved spectroscopy is the collection of techniques that are used to examined the dynamic processes...

Dataset | X-ray Spectroscopy | XFEL

Created 2019/03/15
Size 7 GB
Views 3

jupyterlab

launch VM

Laser-driven Ion Acceleration from Plastic Target

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Fenim ad minim veniam, quis nostrud exerci tationull...

Dataset | Ion Acceleration | ELI Beamlines

Created 2021/11/03
Size 214 GB
Views 7

jupyterlab

launch VM

Electrons accelerated from a thin foil irradiated by an ultra-intense laser

Created 2021/11/03

PaNOSC

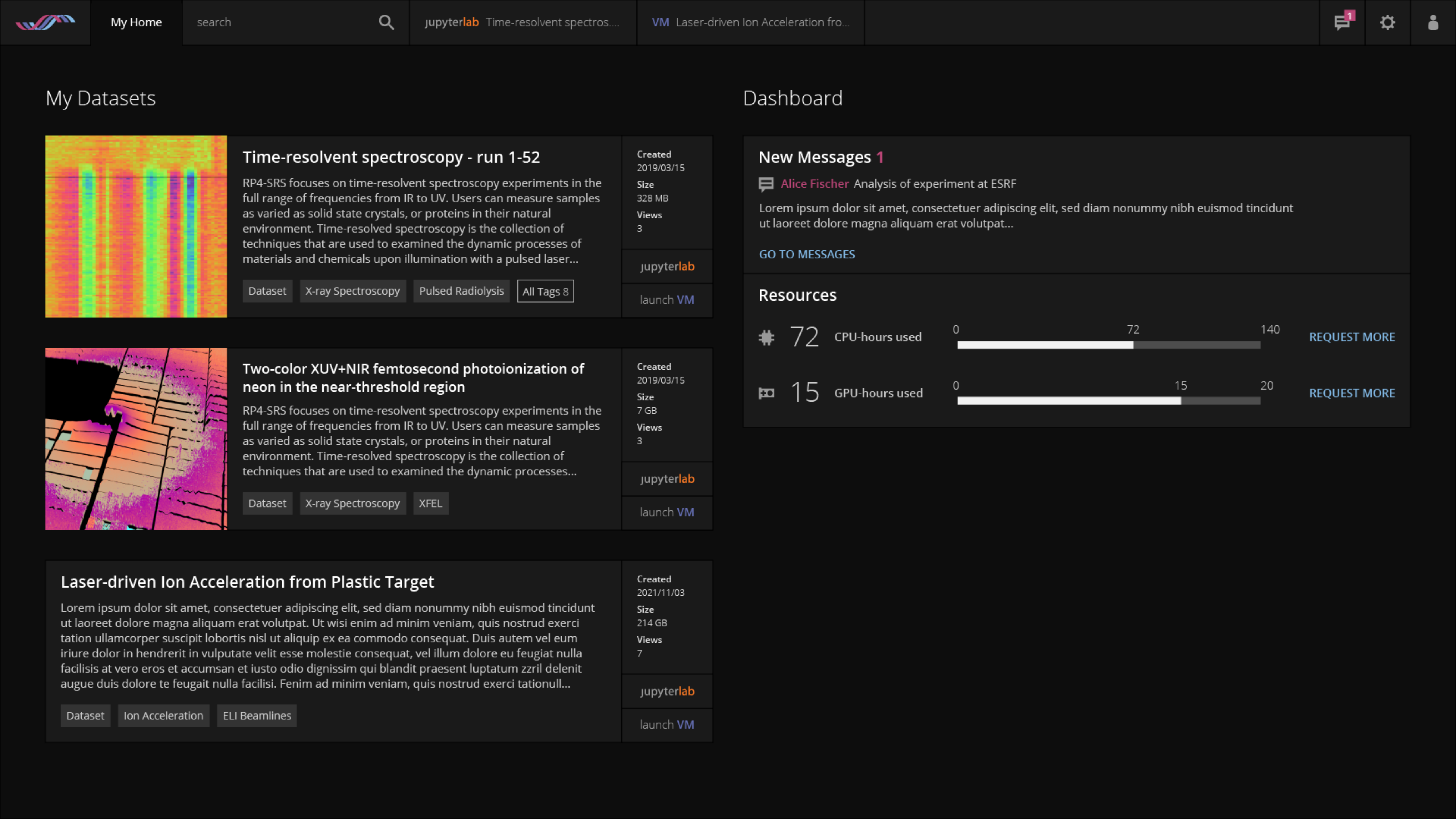
The Photon and Neutron Open Science Cloud (PaNOSC)

The Photon and Neutron Open Science Cloud (PaNOSC) is a European project (financed by the INFRAEOSC-04 call) for making FAIR data a reality in 6 European Research Infrastructures (RIs), developing and providing services for scientific data and connecting these to the European Open Science Cloud (EOSC).

Objectives

- Participate in the construction of the EOSC by linking with the e-infrastructures and other ESFRI clusters.
- Make scientific data produced at Europe's major Photon and Neutron sources fully compatible with the FAIR principles.
- Generalise the adoption of open data policies, standard metadata and data stewardship from 15 photon and neutron RIs and physics institutes across Europe
- Provide innovative data services to the users of these facilities locally and the scientific community at large via the European Open Science Cloud (EOSC).
- Increase the impact of RIs by ensuring data from user experiments can be used beyond the initial scope.
- Share the outcomes with the national RIs who are observers in the proposal and the community at large to promote the adoption of FAIR data principles, data stewardship and the EOSC.

READ MORE


My Home
search
jupyterlab Time-resolvent spectros....
VM Laser-driven Ion Acceleration fro...
1
My Datasets
Time-resolvent spectroscopy - run 1-52
RP4-SRS focuses on time-resolvent spectroscopy experiments in the full range of frequencies from IR to UV. Users can measure samples as varied as solid state crystals, or proteins in their natural environment. Time-resolved spectroscopy is the collection of techniques that are used to examined the dynamic processes of materials and chemicals upon illumination with a pulsed laser...
Dataset
X-ray Spectroscopy
Pulsed Radiolysis
All Tags 8
Created
2019/03/15
Size
328 MB
Views
3
jupyterlab
launch VM
Two-color XUV+NIR femtosecond photoionization of neon in the near-threshold region
RP4-SRS focuses on time-resolvent spectroscopy experiments in the full range of frequencies from IR to UV. Users can measure samples as varied as solid state crystals, or proteins in their natural environment. Time-resolved spectroscopy is the collection of techniques that are used to examined the dynamic processes...
Dataset
X-ray Spectroscopy
XFEL
Created
2019/03/15
Size
7 GB
Views
3
jupyterlab
launch VM
Laser-driven Ion Acceleration from Plastic Target
Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Fenim ad minim veniam, quis nostrud exerci tationull...
Dataset
Ion Acceleration
ELI Beamlines
Created
2021/11/03
Size
214 GB
Views
7
jupyterlab
launch VM
Dashboard
New Messages 1
Alice Fischer Analysis of experiment at ESRF
Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat...
GO TO MESSAGES
Resources
72 CPU-hours used
0
72
140
REQUEST MORE
15 GPU-hours used
0
15
20
REQUEST MORE

Summary

- Introduction PaNOSC project (Photon and Neutron Open Science Cloud), http://panosc-eu.github.io
- Focus on data analysis
- Use cases
  - Remote analysis for huge datasets
  - Make publications reproducible and extensible
  - Allow convenient exploration of existing data sets
- Contributions / brain storming / collaboration welcome

Contact presenter:
robert.rosca@xfel.eu

Contact PaNOSC project:
Andy.Gotz@esrf.fr

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823852.