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
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
    - $\rightarrow$ 2 byte * 1 000 000 * 27 000 / s = 54 GB/s
    - $\rightarrow$ 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
### Other Facilities

<table>
<thead>
<tr>
<th>Data / yr</th>
<th>ILL</th>
<th>ESRF</th>
<th>CERIC</th>
<th>XFEL</th>
<th>ELI</th>
<th>ESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019</td>
<td>200 TB</td>
<td>8 PB</td>
<td>1 PB</td>
<td>3 PB</td>
<td>&lt; 1 PB</td>
<td>0</td>
</tr>
<tr>
<td>2023</td>
<td>600 TB</td>
<td>50 PB</td>
<td>15 PB</td>
<td>100 PB</td>
<td>10 PB</td>
<td>&lt; 1 PB</td>
</tr>
</tbody>
</table>
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)

- 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
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.

Solution architecture for PaNOSC vision:

**Finding data:**
- Web interface with database 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
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 \( \rightarrow \) reproducibility
  - Modify and extend the notebook \( \rightarrow \) reusability

- Users may include scientists, interested public, journal editors and reviewers, representatives from research councils, . . .
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
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
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
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
Jupyter for Reproducible Science at PaN Facilities

Robert Rosca, Reading, 15th October 2019

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

Datasets

Time-resolved spectroscopy - run 1-52

RPM-SRS focuses on time-resolved 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 examine the dynamic processes of materials and chemicals upon illumination with a pulsed laser...

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

RPM-SRS focuses on time-resolved 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 examine the dynamic processes....

Laser-driven Ion Acceleration from Plastic Target

Lorum ipsum dolor sit amet, consectetur adipis sce elit, sed diam nonummy nibh eugum tincidunt ut lectore dolore magna aliquam erat voluput at. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla facilis. At vero eos et accusamus et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore feugait nulla facilisi. Terdint ad minim veniam, quis nostrud exercit

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

The Photon and Neutron Open Science Cloud (PaNOSC)

The Photon and Neutron Open Science Cloud (PaNOSC) is a European project funded 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
Jupyter for Reproducible Science at PaN Facilities

Robert Rosca, Reading, 15th October 2019

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

- Introduction PaNOSC project (Photon and Neutron Open Science Cloud), [http://panosc-eu.github.io](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