SRI2O24

Contribution ID: **785**   Type: **Invited talk**

Managing and Curating Data Flows at PETRA IV and III

Thursday 29 August 2024 16:55 (20 minutes)

The upgrade of the existing PETRA-III synchrotron at DESY to a fourth-generation light source, PETRA-IV, includes not only an increase in brightness but also a new and expanded portfolio of instruments and an updated business model delivering results to non-expert users more promptly than today. Indeed, this business model is already under development at select PETRA-III instruments.

The expectation that a majority of users will be experts in their own scientific fields but not necessarily experts in photon science data analysis highlights the need for the provision of high-level integrated data analysis and data management services to users. We envisage the provision of analytic services to the wider scientific community wherein the timely provision of analysed data to users at conclusion of the measurement is essential. With data volumes exceeding the logistical capacity of most users, and especially non-expert users, these services must be provided by the facility or similar large scale research infrastructure. Provision must also be made for commercial measurement services on top of the same core infrastructure where data must be treated in confidence rather than being destined for open publication.

Prompt analysis of data will become a practical necessity. In the near future a good fraction of the planned instruments will generate in excess of a petabyte of data per day during routine operation, a data volume that already occurs today at some select instruments. Retaining all data on disk for 6 months and on tape for 10 years is no longer economically feasible. Instead, rapid analysis using validated pipelines is required to reduce archived data volumes while providing faster turnaround of results to users performing routine measurements.

Integrated data analysis and data management services are required at the facility to support the full data life cycle from proposal through data taking and on to data analysis, publication, archiving. This includes integrated collection of metadata such as persistent sample identifiers alongside the data, through to infrastructure for enabling data to be open for re-use by the wider community according to FAIR principles (FAIR data stands for Findable, Accessible, Interoperable and Reusable data).

I plan to submit also conference proceedings

Yes

Primary author: BARTY, Anton (FS-SC (Scientific computing))

Presenter: BARTY, Anton (FS-SC (Scientific computing))

Session Classification: Mikrosymposium 3/2: Data, Automation and the Use of AI

Track Classification: 3. Data, Automation and the Use of AI