Contribution ID: **100** Type: **Oral Presentation**

High-Quality AI4S-Oriented Scientific Data Ecosystem for Large Scientific Facilities

Thursday, 19 March 2026 12:06 (22 minutes)

Large Scientific Facilities such as synchrotron radiation facility (e.g., BSRF, HEPS) and spallation neutron sources (e.g., CSNS), are generating massive, complex, and heterogeneous datasets continuously during routine operations and scientific experiments. Managing and utilizing the diverse experimental data, along with simulation results and literature-derived information, is presenting a critical challenge not only in cutting-edge scientific research, and also for industrial applications in the current AI-driven era. We designed and developed a comprehensive data ecosystem aimed at maximizing the scientific data utilization from such facilities, automatically generating high-quality and AI4S-oriented scientific datasets through AI technologies.
First, this paper presents a data ecosystem designed to enhance data quality, accessibility, and reusability, with the ultimate goal of generating high-quality scientific datasets and database. The data ecosystem comprises three core components: data policies & standards, data software & tools, data fusion & provision. Data policies and standards form the foundational element, establishing top-level guidelines and operational frameworks for data collection, storage, sharing, and management. They ensure that data generated across different facilities are interoperable and reusable, enabling cross-facility data sharing. Data software and tools support the entire lifecycle of scientific data, spanning acquisition, management, processing, and analysis. These tools help maintain data quality right from the source of data acquisition. Data fusion and provision utilize data AI agents to align experimental and simulated datasets according to specific research objectives, producing AI-ready datasets. This component also offers versatile APIs and interfaces, facilitating flexible and efficient data access and utilization.
Second, we will report the progress in each component of the ecosystem. In the area of data policies and standards, we are collaborating with multiple scientific facilities to develop a unified data policy and metadata standard, with the goal of establishing it as a national standard (GB) in China. Efforts are also underway to promote and deploy this metadata standard across facilities. For data software and tools, we have developed a suite of frameworks including Mamba for data acquisition, DOMAS for data management, and DAISY for data processing. These tools are tightly integrated with the metadata standards to ensure normative and consistent data handling throughout the entire data lifecycle. In data fusion and provision, data agents have been developed to automate data cleaning, fusion, and alignment. For instance, our synchrotron radiation X-ray diffraction data agent can simulate diffraction data from crystal structure files, automatically process experimental data, and further perform intelligent refinement and deep integration of simulated and experimental data. This process enables the generation of aligned, fused datasets that are AI-ready for model training.
Finally, we summarize the key contributions and outcomes of this work. The proposed data ecosystem significantly improves data management and utilization efficiency, providing a sustainable supply of high-quality datasets that support both the AI4S research paradigm and facility-driven scientific innovation.

Primary author: HU, Hao (Institute of High Energy Physics)

Presenter: HU, Hao (Institute of High Energy Physics)

Session Classification: Artificial Intelligence (AI) - III

Track Classification: Track 10: Artificial Intelligence (AI)