Open Science with openPMD

A. Huebl 1, R. Lehe 1, J.-L. Vay 1, D.P. Grote 1, Ivo F. Sbalzarini 2,5, S. Kuschel 1, M. Bussmann 1

1 Helmholtz-Zentrum Dresden - Rossendorf  2 Technische Universität Dresden
3 Lawrence Berkeley National Lab  4 Lawrence Livermore National Lab
5 Max Planck Institute of Molecular Cell Biology and Genetics  6 Institute for Optics and Quantum Electronics Jena

www.openPMD.org

github.com/openPMD

Self-Description is a Challenge

Scientific workflows need to bridge various applications and algorithms, ideally both automatically and human-readable.

Our glue, using a hierarchical file format such as HDF5, ADIOS BP, XML, JSON, is not automatically scientifically self-describing.

minimal set/kernel of meta information

- meta-standard: truly self-describe data (sinks & sources)
- open-access: unified description (creation → publishing)
- workflows: high-level integrations (apps, visualization markups)

Key Concepts by Example

A strict grouping but flexible naming of records allows easy parsing and traversal.

Heavy data is guaranteed to stay contiguous for performant I/O. Light-weight annotations are buffered and read/written at once.

Example for the structure of an openPMD annotated data set. From a user-point of view, records are the central objects to be described.

Exascale Computing Needs Multi-PByte Scalable, Documented Data

User-space expressive:

- constant record components
- domain patches

Still full functionality of underlying I/O libraries:

- portability
- internal / external links
- strides, aggregations, multi-file
- compression [2], staging [3,4]

Integrated and long staged I/O pipelines will be essential for I/O in Exascale HPC. Meta-data must easily propagate and be usable at any stage and time.

Open Science Attracts Collaboration

- source: open, contributable
- review: open issues/updates
- methodology: documented workflows
- education: resources & integrations
- data: versioned, self-describing

reproducibility
quality
sustainability
exchange
after-use

Open Simulations:

- PIConGPU [HDR], ParaTAXIS [HDR],
- openFPM [MPI, CBG], Warp [LNL, LNL],
- FBPIC [LNL, DESY], SIMEX [CALC]

Open Post-Processing:

- openPMD-viewer + contribute & reuse: yt project,
  Visit, postpic, pyDive, XDMF, HDF Compass, libSshlash, ...

PByte-Scale: PIConGPU I/O on Titan [2]

http://www.openPMD.org


This research used resources of the Oak Ridge Leadership Computing Facility located in the Oak Ridge National Laboratory, which is supported by the Office of Science of the Department of Energy under Contract DE-AC05-00OR22725. Supported in part by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Prepared in part by LNL under Contract DE-AC02-07NA27344. The authors are thankful for the community contributions to libraries, software ecosystem, user support, review and integrations. Particularly, thank you to Yaser Afshar, Richard Briggs, Heiko Burau, Jong Choe, Marcia Garten, Daniel Grassinger, Alexander Grund, Carsten Fortmann-Grote, Sören Jalas, Manuel Kirchen, Scott Klasky, Noah Klemm, Fabian Koller, Mathieu Lobet, Richard Pausch, Norbert Podhorszki, David Pugmire, Felix Schmitt, Klaus Steininger, Michael Sippel and René Widera!