Open Science with openPMD

A. Huebl[1,2], R. Lehe[3], J.-L. Vay[3], D.P. Grote[4], Ivo F. Sbalzarini[2,5], S. Kuschel[6], 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

the **meta-data standard**

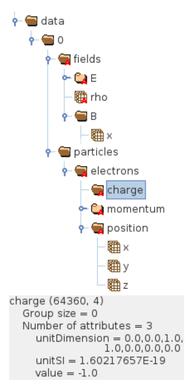
- **p**article and **m**esh based **d**ata
- **data format agnostic**
- **frictionless data exchange**

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

electric field $\vec{E}(\vec{r})$: ← **record**

/ ... / meshes / E / x, y, z ← **component**

.unitSI, .unitDimension,

.geometry, .time, ... ← **attribute**

electron charge Q_i: ← **group**

/ ... / particles / electrons / Q_i

.unitSI, ...

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

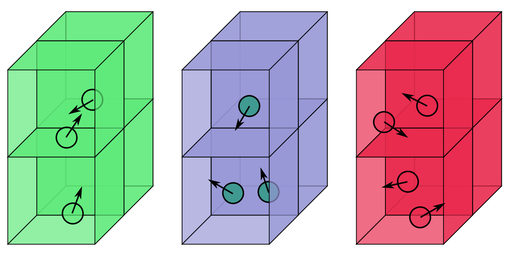
Heavy data is guaranteed to stay contiguous for performant I/O. Light-weight annotations are buffered and read/written at once.

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 expressible:

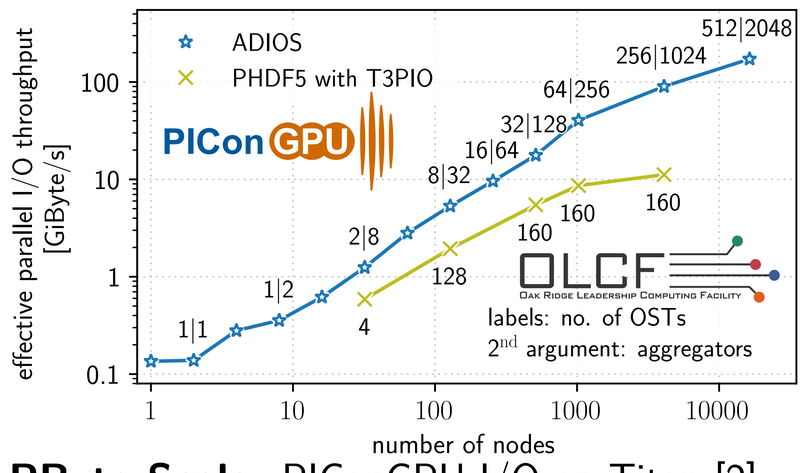
- constant record components
- domain patches



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.



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

Open Science Attracts Collaboration

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

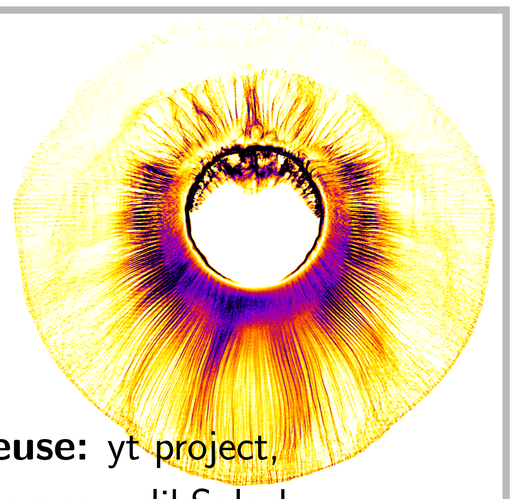
→ **reproducibility**, **quality**, **sustainability**, **exchange**, **after-use**

Open Simulations:

PIConGPU[HZDR], ParaTAXIS[HZDR], openFPM[MPI-CBG], Warp[LBNL, LLNL], FBPIC[LBNL, DESY], SIMEX[EUCALL]

Open Post-Processing:

openPMD-viewer + **contribute & reuse:** yt project, VisIt, postpic, pyDive, XDMF, HDF Compass, libSplash, ...



[1] A. Huebl et al. *openPMD 1.0.0: A meta data standard for particle and mesh based data*, technical specification (CC-BY 4.0), November 2015, DOI:10.5281/zenodo.33624 [2] A. Huebl et al. *On the Scalability of Data Reduction Techniques in Current and Upcoming HPC Systems from an Application Perspective*, ISC 2017, arXiv:1706.00522 [3] H. Abbasi et al. *Datastager: scalable data staging services for petascale applications*, Cluster Computing 13(3), DOI:10.1007/s10586-010-0135-6 [4] C. Docan et al. *DataSpaces: An interaction and coordination framework or coupled simulation workflows*, HPDC 2010, DOI:10.1007/s10586-011-0162-y

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 LLNL under Contract DE-AC52-07NA27344. The authors are thankful for the community contributions to libraries, software eco-system, user support, review and integrations. Particularly, thank you to Yaser Afshar, Richard Briggs, Heiko Burau, Jong Choi, Marco 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 Steiniger, Michael Sippel and René Widera!

LLNL-POST-733711

DOI:10.5281/zenodo.822396

This project has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No 654220.