Develop/integrate data standards & start-to-end workflows for Accelerator Physics

(Letter of Interest to Snowmass21, Computational & Accelerator Frontiers)

A. Huebl*[1], J.-L. Vay[1], R. Lehe[1], M. Thévenet[2], C. Mayes[3], D. Sagan[4], Y.-D. Tsai[5], J. C. E[6], F. Tsung[7], H. Vincenti[8], A. Ferran Pousa[2], N. M. Cook[9], S. J. Gessner[3], F. Poeschel[10], M. Bussmann[11,10], D. P. Grote[12], N. A. Murphy[13], R. Schmitz[14], C. H. Yoon[3], D. L. Bruhwiler[9], K. Cranmer[15], S. R. Yoffe[16], B. Cros[17], A. L. Edelen[3], G. Stark[18]

[1] *Lawrence Berkeley National Laboratory, Berkeley, California 94720 USA*
[2] *Deutsches Elektronen Synchrotron (DESY), Hamburg, Hamburg 22607 Germany*
[3] *SLAC National Accelerator Laboratory, Menlo Park, California 94025 USA*
[4] *Cornell University, Ithaca, New York 14850 USA*
[5] *Fermi National Accelerator Laboratory (Fermilab), Batavia, Illinois 60510 USA*
[6] *European XFEL GmbH, Schenefeld, Schleswig-Holstein 22869 Germany*
[7] *University of California, Los Angeles, CA 90095 USA*
[8] *LIDYL, CEA-Université Paris-Saclay, CEA Saclay, 91 191 Gif-sur-Yvette, France*
[9] *RadiaSoft LLC, Boulder, Colorado 80301 USA*
[10] *Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Dresden, Saxony 01328 Germany*
[11] *Center for Advanced Systems Understanding (CASUS), Görlitz, Saxony 02826 Germany*
[12] *Lawrence Livermore National Laboratory, Livermore, California 94550 USA*
[13] *Center for Astrophysics | Harvard & Smithsonian, Cambridge, Massachusetts 02138 USA*
[14] *University of California, Santa Barbara, California 93106 USA*
[15] *New York University, New York, NY 10003 USA*
[16] *SUPA and University of Strathclyde, Glasgow G4 0NG, United Kingdom*
[17] *CNRS, Université Paris Saclay, Orsay, 91400 France*
[18] *SCIPP, UC Santa Cruz*

August 2020

Abstract

*axelhuebl@lbl.gov

1 Topic and Status

Research roadmaps in accelerator and beam physics are multi-decade activities, designing machines from source to final detector and whose realizations are funded with billions of dollars. The data produced in both experiments as well as modeling efforts is the basis for an evidence-based research approach.

Due to the high cost of these activities, detailed modeling activities are essential before investing in upgrades and new constructions, and are instrumental in optimizing existing facilities. It arises naturally that the expertise in the field could be brought together in the development of virtual twins of facilities [1], combining the know-how of the community and producing research data curated in a science gateway. As of today, a variety of software tools are deployed for design, pre-processing, modeling and analysis of the specifications in current and future machines. Many of these tools are developed by various groups and developers with varying focus and conventions. Consequently, physical input descriptions are often specific to existing modeling tools, which complicates data exchange, testing new theories against large sets of previous results, and integration efforts such as start-to-end modeling that could advance the field's predictive capabilities.

As of today, except for activities such as those supported by the Consortium for Advanced Modeling of Particle Accelerators (CAMPA), there is little coordination between different groups and activities that address how to integrate existing frameworks into newly emerging workflows, such as complex multi-physics scenarios between several simulation frameworks. This essentially leads to underutilized opportunities, re-invention and under-maintained research tools, isolated solutions, longer integration cycles when exploring new approaches, and consequently slower scientific process. This situation is at least in part reinforced by challenges integrating computational science methods in modern physics curricula [2], need for new cross-domain career paths [3], and sometimes limited adoption of open science principles [4].

2 Current and future challenges

2.1 Data

There exist a variety of low-level data file formats, which individual modeling tools use with custom-made meta-data conventions to express the domain-specific data consumed and generated by individual tools. New data formats and highly-tuned I/O libraries are continuously being developed by computer scientists and existing paradigms, such as POSIX-I/O, are overtaken by modern approaches such as data streams, object storage and relaxed constrains in highly-parallel I/O. Flexibility is needed with respect to these low-level file formats, to quickly utilize progress in modern storage and data transport technology.

Furthermore, contemporary and planned advances in high-repetition rate experiments as well as simulation needs for large-scale, high-resolution modeling and high-throughput computing continue to increase data rates and data size demands significantly (by orders of magnitude). As an example, advanced accelerators anticipate kHz facilities and simulation requirements are still making full use of the exponentially growing computing capabilities at top-tier supercomputing centers - notably with significantly slower growth in available filesystem storage bandwidth. Under these developments, file-based storage of high-fidelity data as well as manual data analysis and curation efforts are increasingly becoming the bottleneck in the scientific process.

2.2 Workflows

Seldom is a single code sufficiently complete, accurate, and fast to simulate the full range of a physics design without introducing many approximations. Most tools used for accelerator and beam modeling are not compatible nor developed with the goal to be reused in overarching workflow [5]. Workflows are often manual and mostly rely on access to domain-specific raw data, which may not be integrated into science gateways for curation and sharing.

Productive reproducibility, such as re-playing the simulation design of another group, should become the default for validating new physical results that increasingly rely on computational backing. This needs a high degree of documentation and automation, overcoming manual and person-hour intensive sample comparisons.

Workflow tools that address the needs in a general way are emerging, rapidly evolving and need effort (integration and continuous maintenance in an evolving software ecosystem) to be exploited for accelerator and beam modeling. Generalizing design descriptions (inputs) to be shared between codes require the same level of long-term commitment.

3 Advances needed to meet challenges

Generally, many in the community agree to address those challenges with standards that everyone can use, integrate and contribute to [6]. With active contributions, standards are not set in stone but can be developed just like software in versions as arising challenges are adopted and agreed upon, limitations are identified, and refined solutions are emerging. Standards can emerge from a good documentation that is refined by including a wider audience into the process, evolving into transparent decision processes and adoption of best practices over time (versioning, citations for documents/data/software, compatibility strategies, open community governance, etc.).

3.1 Data

For data exchange, defining meta-data in a file-format agnostic organization has been successfully demonstrated for accelerator and beam data in the Open Standard for Particle-Mesh Data Files (openPMD) [7].[1] openPMD organizes around a written, versioned text document that is supported by tooling for validation, examples, a project catalogue, libraries and programs, which all have their respective documentation and tutorials. Individual, compatible projects, software and data are published by a variety of authors [8].

The Standard Input Format for Particle-In-Cell Codes (PICMI) proposed to address the challenge of simulation design sharing by defining a common input layer that focuses on the physical description of a problem set [9]. Currently implemented as a Python API description, simulations can be designed flexibly with a well-known programming language in the community.[2]

Standardization of inputs and outputs could also provide the basis for an integration into scientific data portals to curate and re-use modeling data [10, 11, 12]: this can be raw and/or derived (output) data with respective meta-data describing the physical scenario (input) and aide meta-studies and tests for new theories.

After successful initial integrations, ongoing activities need to aim for widened support in newly emerging, complex analysis tools (from desktop to supercomputer). Also, code-coupling workflows are based on data exchange and could benefit in many cases from data streaming, a technique that is used in tokamak modeling (XGC/Gene) as well as particle-in-cell modeling (PIConGPU/GAPD). With a well-defined domain-specific standardization, efforts of few, specialized people in the community are needed for adopting modern, low-level computer science data libraries as they emerge. In parallel, modern educational materials are needed that demonstrate and introduce such scientifically self-describing I/O to newly joining community members.

3.2 Workflows

One needed advance that connects scientific data and workflows is the need to be able to reproduce start-to-end workflows. For example, many traditional file-based workflows need to be adopted to in situ workflows in order to bridge data production rate and storage capability gaps. But a simulation pipeline with in situ data analysis components needs to be flexible for adopting use cases and re-playable from the start, if raw data exceeds the long-term storage capabilities. Consequently, the software and data roadmap strategies in the community need to be as long-term as the machines and experiments which they model.

Ideally, workflows and components could be based on compatible, modular toolkits and libraries [5, 13]. Similar advances as start-to-end modeling workflows that have been demonstrated for XFEL facilities (e.g. PaNOSC [14], SIMEX [15], LUME [16]) can be made, e.g. for plasma-based accelerators. As in the mentioned examples, domain-agnostic workflow tools such as Fireworks [17], SnakeMake, Flux [18] are readily available to implement workflows. As many workflows will aim for optimization and exploration [19, 20], managing many modeling runs with frameworks [21, 22, 23], especially in the context of machine learning algorithms [24], will become essential. Again, standardization of workflows (interfaces, APIs, compatible modules, data) can lower the entry burden significantly, ensure a close connection to long-term data curation, and foresee extensibility of workflows over time.

[1] Similarly, event and detector data is successfully defined in packages of ROOT that is used by many independent software.

[2] Projects that define a common API, e.g. BLAS for linear algebra and MPI for multi-node message passing, use a similar approach.

References

[1] Jean-Luc Vay et al. "EVA (End-to-end Virtual Accelerators)". In: *Snowmass21 LOI* (2020).

[2] Ryan Schmitz, Tom Eichlersmith, and Axel Huebl. "Barriers to Entry in Physics Computing: A Snowmass Letter of Interest for the Computational Frontier". In: *Snowmass21 LOI* (2020).

[3] Snowmass Early Career (Accelerator Frontier): Nathan Cook et al. "Set for Success: How to Accelerate Early Career Scientists". In: *Snowmass21 LOI* (2020).

[4] Axel Huebl et al. "Aspiration for Open Science in Accelerator & Beam Physics Modeling". In: *Snowmass21 LOI* (2020).

[5] Jean-Luc Vay et al. "A modular community ecosystem for multiphysics particle accelerator modeling and design". In: *Snowmass21 LOI* (2020).

[6] Jean-Luc Vay et al. "Integrated ecosystem of advanced simulation tools for plasma modeling". In: *White paper for the 2020 NAS Decadal Study on Plasma Science* (2019), submission no. 76. URL: https://app.smartsheet.com/b/publish?EQBCT=40f1147f433f4e858312ba75af14d70f.

[7] Axel Huebl et al. "openPMD: A meta data standard for particle and mesh based data". In: (2015). DOI: 10.5281/zenodo.591699. URL: https://doi.org/10.5281/zenodo.591699.

[8] *Curated catalogue of projects supporting openPMD*. URL: https://github.com/openPMD/openPMD-projects.

[9] *PICMI: Standard input format for Particle-In-Cell codes*. URL: https://github.com/picmi-standard.

[10] *CERN Data Portal*. URL: http://opendata.cern.ch.

[11] *FAIR Principles*. URL: https://www.go-fair.org/fair-principles/.

[12] *RECAST framework reinterpretation of an ATLAS Dark Matter Search constraining a model of a dark Higgs boson decaying to two b-quarks*. Tech. rep. ATL-PHYS-PUB-2019-032. Geneva: CERN, Aug. 2019. URL: https://cds.cern.ch/record/2686290.

[13] David Sagan et al. "Beam Dynamics Toolkit". In: *Snowmass21 LOI* (2020).

[14] *The Photon and Neutron Open Science Cloud (PaNOSC)*. en-GB. URL: https://www.panosc.eu/.

[15] *PaNOSC-ViNYL/SimEx*. May 2020. URL: https://github.com/PaNOSC-ViNYL/SimEx (visited on 07/19/2020).

[16] *Lightsource Unified Modeling Environment (LUME)*. URL: https://github.com/slaclab/lume.

[17] Anubhav Jain et al. "FireWorks: a dynamic workflow system designed for high-throughput applications". In: *Concurrency and Computation: Practice and Experience* 27.17 (2015), pp. 5037–5059. DOI: 10.1002/cpe.3505. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.3505.

[18] Dong H. Ahn et al. ""Flux: Overcoming scheduling challenges for exascale workflows"". In: *Future Generation Computer Systems* 110 (2020), pp. 202–213. ISSN: 0167-739X. DOI: https://doi.org/10.1016/j.future.2020.04.006. URL: http://www.sciencedirect.com/science/article/pii/S0167739X19317169.

[19] Qianqian Su et al. "Optimization of beam qualities on Plasma Wakefield Acceleration". In: *APS Division of Plasma Physics Meeting Abstracts*. Vol. 2019. APS Meeting Abstracts. Jan. 2019, GO5.006.

[20] Remi Lehe. *Using Gaussian Processes to target the injection boundary in laser-plasma simulations*. 2nd ICFA Mini-Workshop on Machine Learning for Charged Particle Accelerators. 2019. URL: https://indico.psi.ch/event/6698/contributions/16533/.

[21] Stephen Hudson et al. *libEnsemble Users Manual*. Tech. rep. Revision 0.7.0. Argonne National Laboratory, 2020. URL: https://buildmedia.readthedocs.org/media/pdf/libensemble/latest/libensemble.pdf.

[22] *OCELOT Generic Optimizer for accelerators*. https://github.com/ocelot-collab/optimizer.

[23] Christopher Mayes. *xopt: Simulation optimization, based on DEAP*. https://github.com/ChristopherMayes/xopt.

[24] Remi Lehe et al. "Machine learning and surrogates models for simulation-based optimization of accelerator design". In: *Snowmass21 LOI* (2020).