Summary WG3
Data Generation & Simulations

WG Members

19 members present on workshop:

- Adrian Oeftiger (GSI/FAIR)
- Adnan Ghribi (CEA/CNRS)
- Amelia Pollard (ASTEC/STFC)
- Andrea Santamaria Garcia (KIT)
- Barbara Dalena (CEA/IRFU)
- Chenran Xu (KIT)
- Damien Minenna (CNRS)
- Davide Di Croce (EPFL)
- Francesco Massimo (UPS)
- Francis Osswald (CNRS/IPHC)
- Gianluca Valentino (UM)
- Hayg Guler (CNRS/IJCLab)
- Kevin Cassou (CNRS)
- Mads Bertelsen (ESS)
- Marcel Heine (CNRS/IPHC)
- Pierre Schnizer (HZB)
- Simon Hirlaender (PLUS)
- Simone Liuzzo (ESRF)
- Thomas Kachelhoffer (CCIN2P3)

Topics covered:

- Accelerator physics
- Accelerator technology
- Linacs / Rings
- Hadron beams / light sources
- Plasma WF accelerators
- Neutron scattering

DA and loss rate predictions and FCC optimisation (Davide Di Croce, Tatiana Pieloni, Ekaterina Krymova, et al)

DA and collective effects (Adnan Ghribi, Barbara Dalena, Quentin Bruant)

Exploring Resonance Diagrams (Adrian Oeftiger et al)

HZB / Pierre Schnitzer

Surrogate modeling of beam losses in the LHC collimation hierarchy (Gianluca Valentino)

Online prediction of lifetime / beta-beating / injection eff. for ESRF (Simone Maria Liuzzo)

Machine learning based control for temporal and spatial laser shaping at CLARA (Amelia Pollard)

Inverse Design for Neutron Scattering Experiment (Kim Lefmann)

~ Data classification for background removal (Mads Bertelsen)

Design of a multi-stage laser wakefield accelerator at LPGP (Francesco Massimo)

Enhanced Emittance Evaluation (Marcel Heine, Francis Osswald)

Pool of Study Cases

PIC simulations dataset for laser-plasma injector optimisation (Kevin Cassou)

EARLI, Design of a laser wakefield accelerator for AWAKE (Damien Minenna)

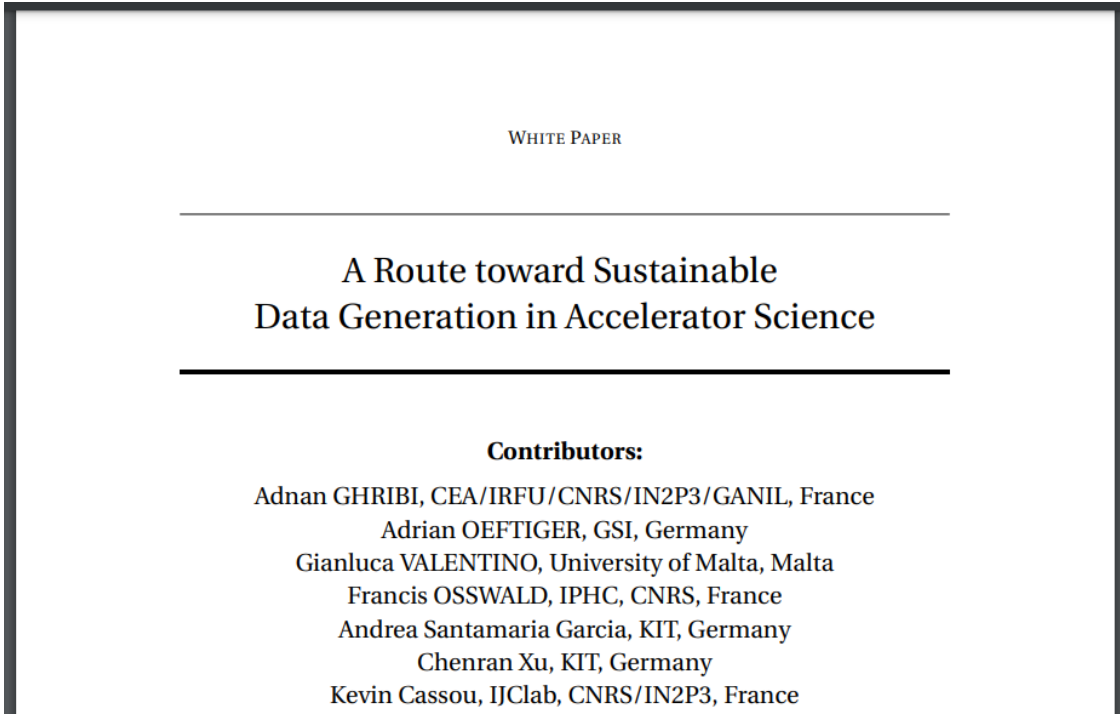
Automated data collection for laser-plasma injector beam quality optimisation (Kevin Cassou)

Surrogate Modelling of a Low-energy Linac (Chenran Xu, Andrea Santamaria Garcia)

Surrogate modeling for ThomX (Hayg Guler)

White Paper to Shape Ideas

- Data: bottom-up strategy for WG
  - Compare data management solutions: existing repos, EOSC projects (e.g. ESCAPE DIOS/OSSR)
  - Build on existing infra-structure and devise data publication workflow
- Pool: 14 study cases (+1)
- Investigate methods for active learning (=iterative supervised learning)
- Finalise & publish ~October ‘23



WHITE PAPER

A Route toward Sustainable Data Generation in Accelerator Science

Contributors:

Adnan GHRIBI, CEA/IRFU/CNRS/IN2P3/GANIL, France
Adrian OEFTIGER, GSI, Germany
Gianluca VALENTINO, University of Malta, Malta
Francis OSSWALD, IPHC, CNRS, France
Andrea Santamaria Garcia, KIT, Germany
Chenran Xu, KIT, Germany
Kevin Cassou, IJClab, CNRS/IN2P3, France

Deliverables

- Standardising data:
  - Data: models & structure
  - Standardisation of metadata
  - Strategise data publication workflow
- Developed translation tool that curates and structures data
  → produce F.A.I.R. data (potentially published with DOI) to enable further data-driven processing
- Generated F.A.I.R. data sets
- Developed framework for simulations:
  - Input generation: clever parameter sampling
    ([uncertainty-aware] active learning algorithms)
    - Improve on grid search & random search!
  - Interpreter: input parameters → simulation codes
  - Include translation tool (simulation code → metadata compatible)
  - potentially: cluster job allocation management
  - potentially: exploit trained surrogate (with uncertainty measure) to re-use data/information

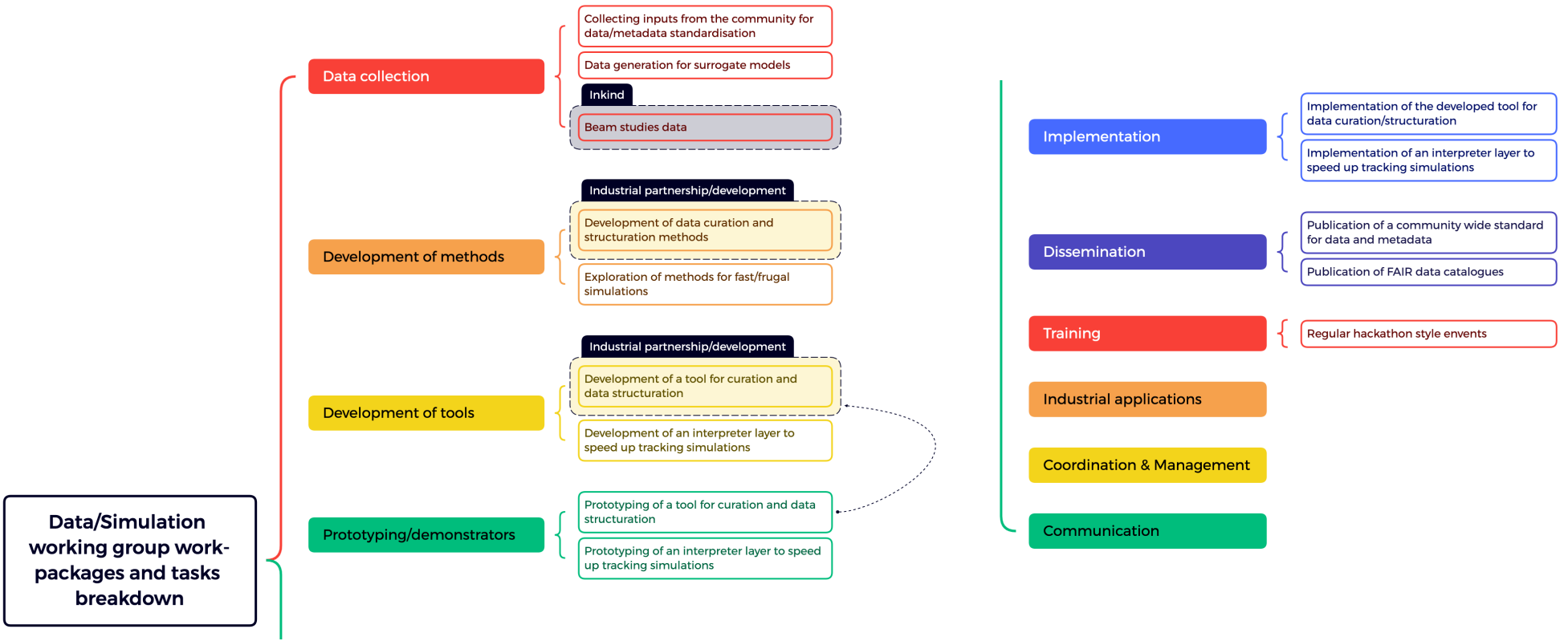
Approach

- Regular hackathon-style events, per event:
  - Prepare 1-3 study cases as boiled down projects
  - Implement active learning strategies,
    compare effectiveness/energy-efficiency
  - Generate data sets,
    compare data management / publication solutions
  - Develop standardised set of metadata

Impact

- #scientific publications / data catalogues
- #codes using the defined data standard
- Reduction of #data & speed-up factor required to reach conclusion in study
- Energy consumption & footprint (storage, computing)
- Geographical spread of (framework/data catalogue) end users
- Training/formation: #people and #data challenges (hackathon events)
- Knowledge transfer & validation across institutes
- Framework applicable to use cases in wider community (extensibility/plugins)

Draft Structure for WP (to be synced)



Resources

- FTE on data/metadata collection
- Long-term support of simulation-embedding framework (in-kind, FTE)
- FTE or industry partnership
  (e.g. software companies on MLops / DEVops)
- Beam time (in-kind)
- Infrastructure for data storage & computing
- Commitment from research infrastructure (data stewards/officers)
- statistics/ML experts (from universities/industry)

Next Steps

- Add workshop outcome to white paper
- Conclude on Tasks & roles of WG members
- Synchronise WP structure with other WGs
- Build industrial partnerships

Backup

Collected Expected Results

- Several methods of active learning / smart data exploration
- Online dynamic exploration of published results
- New tools to design (LWFA and classical accelerators, neutron scatt. exp.) front-to-end
  $\Longrightarrow$ inverse design
- Interfacing tools between codes modelling different physics (at a given fidelity/precision)
- Data sets (to train AI & obtained surrogates during active learning)
- Surrogate models for large machines (present & future) $\rightarrow$ with aim to embed in optimisation/classification problems
- Synthetic data for precise surrogate-optimisation purposes (data augmentation)
- Enhanced diagnostic (background removal)
- Implemented & shared framework managing (i) clever parameter generation, (ii) cluster job allocation, (iii) data postprocessing (implementing agreed data standards)
- Knowledge transfer & validation across institutes
- Metadata catalogue
- Have produced a F.A.I.R. & reused data set (citations of DOI, need to establish “impact factor of shared data”)