Universität Regensburg

Provenance for Lattice QCD Workflows – An Update

Tanja Auge[1], Meike Klettke[1], Gunnar Bali[2], Christian Kindler[2], Daniel Knüttel[2], Wolfgang Söldner[2], Tilo Wettig[2]

[1]University of Regensburg, Faculty of Informatics and Data Science, Germany
[2]University of Regensburg, Department of Physics, Germany

Objective & Contributions

Motivation: missing provenance support for (complex, multi-step, and distributed) Lattice QCD workflows to answer provenance questions like

- *Q1*: Which datasets are affected by an error or bug?
- *Q2*: How are datasets affected by modifying a parameter?
- *Q3*: Who was involved in generating the data?

Aim:

- enable reproducible, traceable data analysis in lattice QCD workflows

Contribution:

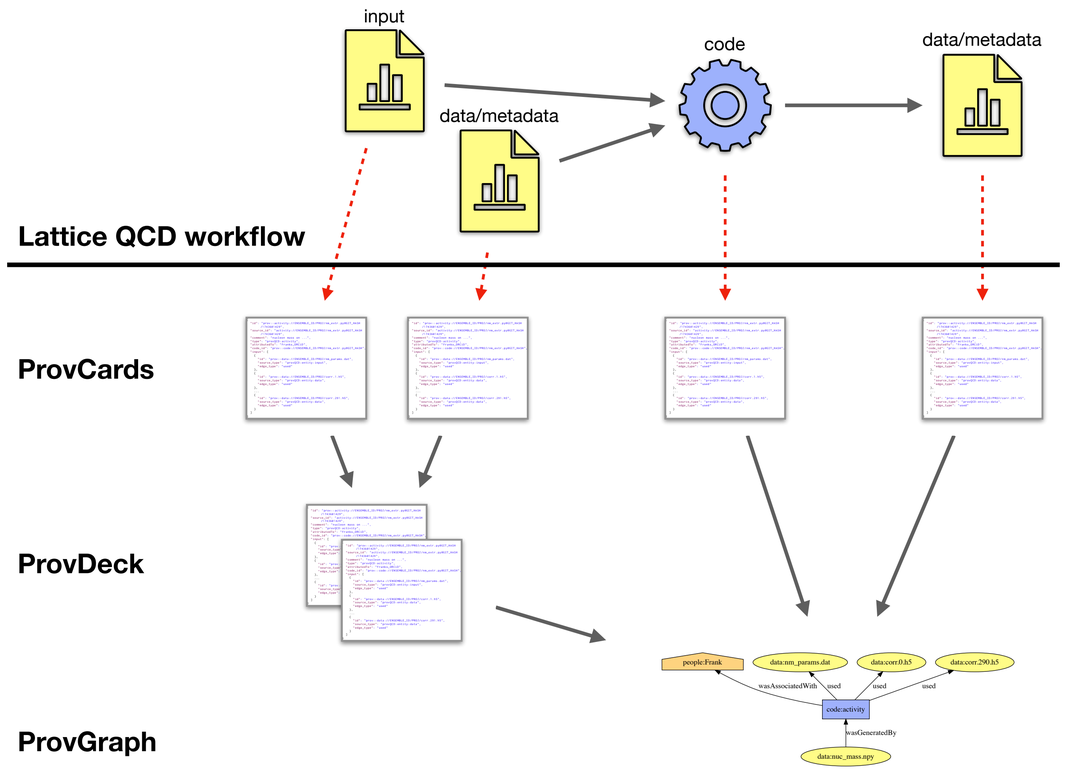
- ProvCards/ProvDecks isolate and organize provenance-related metadata
- ProvGraphs visualize workflow provenance

Provenance for Lattice QCD

- a typical Lattice QCD workflow factorizes into three phases:
  - *generation*: ensembles of gauge-field configurations
  - *measurement*: correlation functions
  - *analysis*: computation of observables
- existing community standards for Lattice QCD gauge-field configurations, *ensemble metadata* and *configuration metadata* [2, 3] ⇒ provenance-related metadata for generation phase
- existing PROV-template for generation and measurement phase [1]

In summary: existing standards and concepts provide support only for the initial two phases.

Provenance Concept

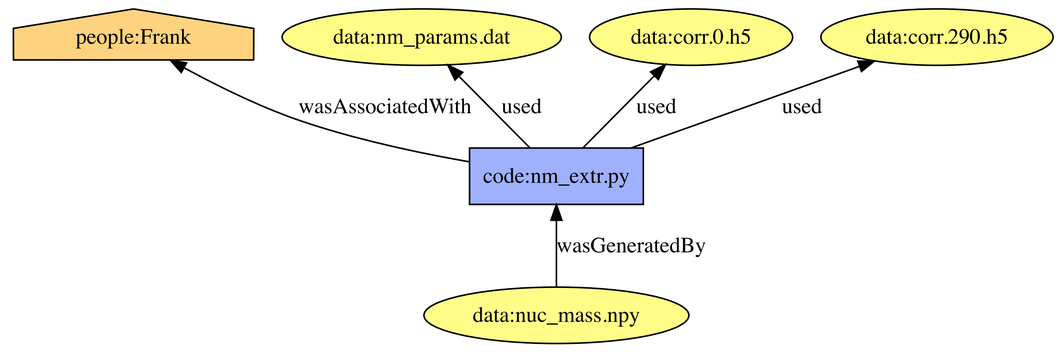


ProvCards, ProvDecks, and ProvGraphs

ProvCard: machine-readable file (e.g., JSON) that contains provenance-related metadata for a specific source (entity, agent, activity)

```
{
  "id": "prov::activity://ENSEMBLE_ID/PROJ/nm_extr.py@GIT_HASH
       /1743601429",
  "source_id": "activity://ENSEMBLE_ID/PROJ/nm_extr.py@GIT_HASH
       /1743601429",
  "comment": "nucleon mass on ...",
  "type": "provQCD:activity",
  "attributedTo": "Franks_ORCiD",
  "code_id": "prov::code://ENSEMBLE_ID/PROJ/nm_extr.py@GIT_HASH",
  "input": [
    {
      "id": "prov::data://ENSEMBLE_ID/PROJ/nm_params.dat",
      "source_type": "provQCD:entity:input",
      "edge_type": "used"
    },
    {
      "id": "prov::data://ENSEMBLE_ID/PROJ/corr.1.h5",
      "source_type": "provQCD:entity:data",
      "edge_type": "used"
    },
    ...
    {
      "id": "prov::data://ENSEMBLE_ID/PROJ/corr.291.h5",
      "source_type": "provQCD:entity:data",
      "edge_type": "used"
    }
  ]
}
```

ProvDeck: collection of ProvCard
ProvGraph: generated provenance graph that visualizes the relationships between entities, agents, and activities according to the W3C PROV-DM



Next steps

- specification of a metadata standard for ProvCards and a corresponding ontology for the Lattice QCD community
- further development of the *provQCD toolbox* for
  - ProvCard extraction from metadata and log files
  - generation of ProvDecks and ProvGraphs from ProvCards

Acknowledgement

This work was supported in part by the Deutsche Forschungsgesellschaft (DFG) under grant PUNCH4NFDI, project number 460248186.

References

[1] T. Auge et al. Provenance for Lattice QCD workflows. In *WWW (Companion Volume)*, pages 1524–1530. ACM, 2023.

[2] ILDG. Specification of ILDG Standards. `https://hpc.desy.de/ildg/specifications/`, 2025.

[3] B. Sagar et al. International Lattice Data Grid 2.0: Status and Progress. *PoS*, LATTICE2024:411, 2025. doi: 10.22323/1.466.0411.