KM3NeT provenance

Status and plans

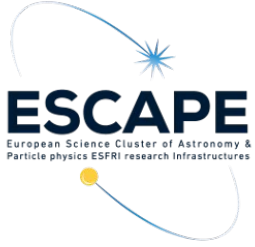
Tamas Gal, Jutta Schnabel

ESCAPE WP4 Provenance Workshop

7th September 2020

ERLANGEN CENTRE
FOR ASTROPARTICLE
PHYSICS

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

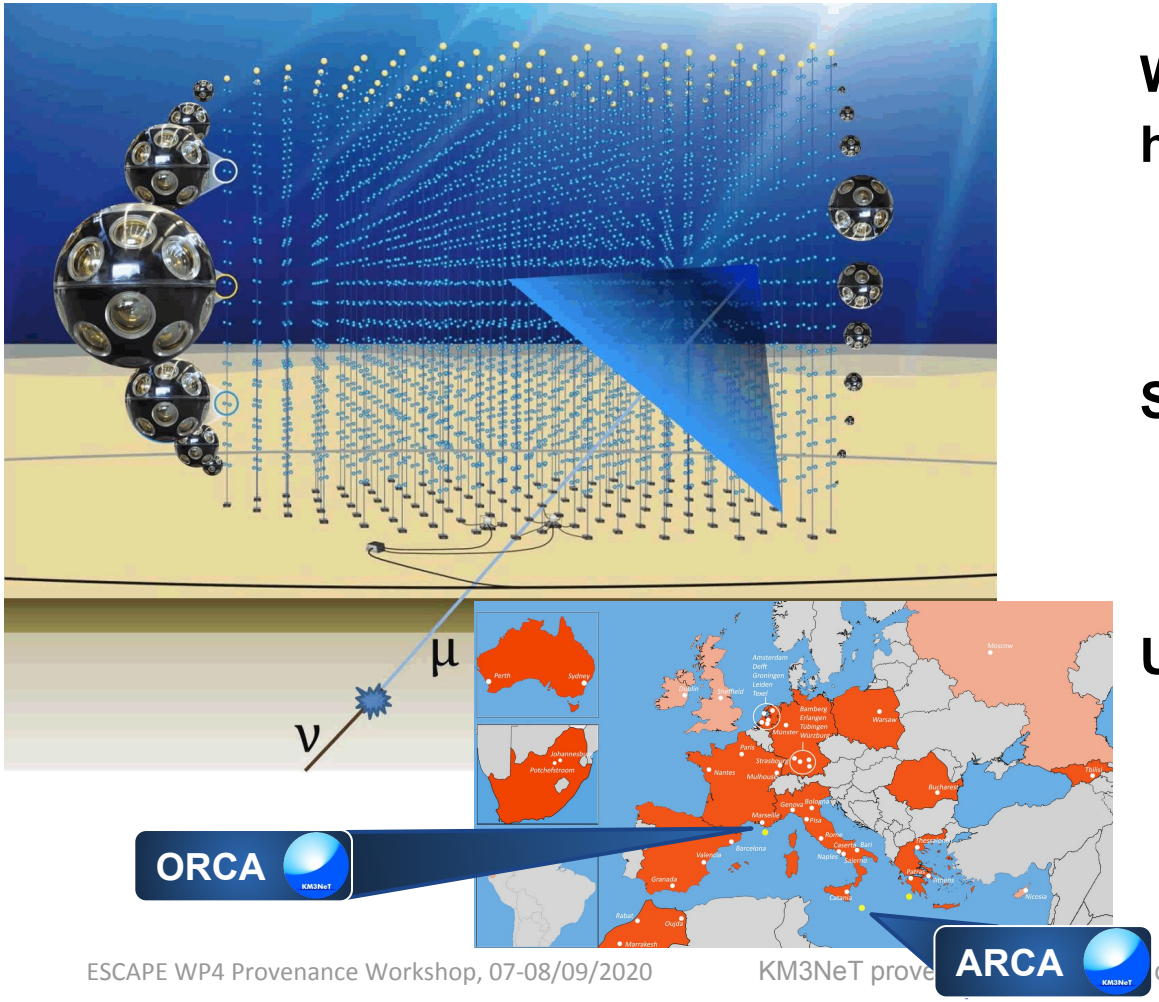


KM3NeT provenance

Disclaimer: concepts discussed here are under construction and discussion, showing the current status only!

- Tier-based processing of particle detections (“events”) with high-volume complex processing workflow on HTC clusters
- Need both detailed *data generation* and *data processing* provenance
- for data processing, extended provenance information beyond simple “in-out tracking” is needed.

KM3NeT experiment



Water Cherenkov detector for high-energy neutrinos

- Multi-PMT sensor modules
- Building blocks (BBs) of 115 DUs (lines)

Science goals

- astrophysics (ARCA)
- neutrino oscillations (ORCA)

Under construction

- 1 DU installed in ARCA
- 6 DUs installed in ORCA
- first stage: 24 ARCA + 6 ORCA strings
- goal: 2BB ARCA + 1 BB ORCA

Open science data products

Data processing

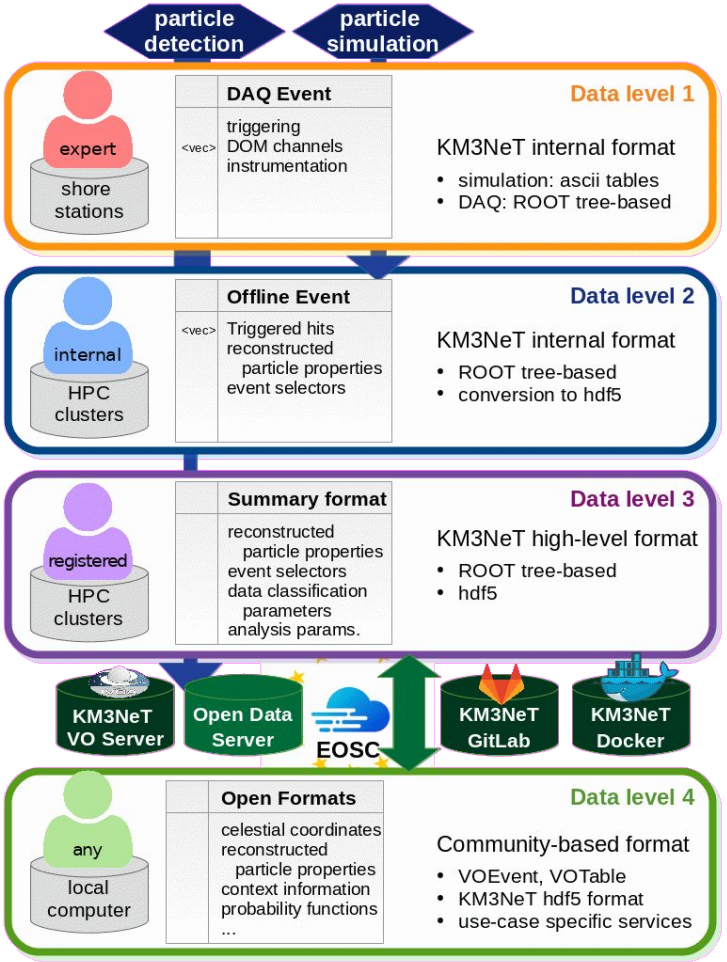
- Typical for particle physics, tier-based
- High-volume data taking
- Triggering
- Processing: reconstruction & filtering

Event simulation

- Signal and background events
- Analogous processing to measurements

→ need of distributed PB-sized storage
and 10k+ CPU resources

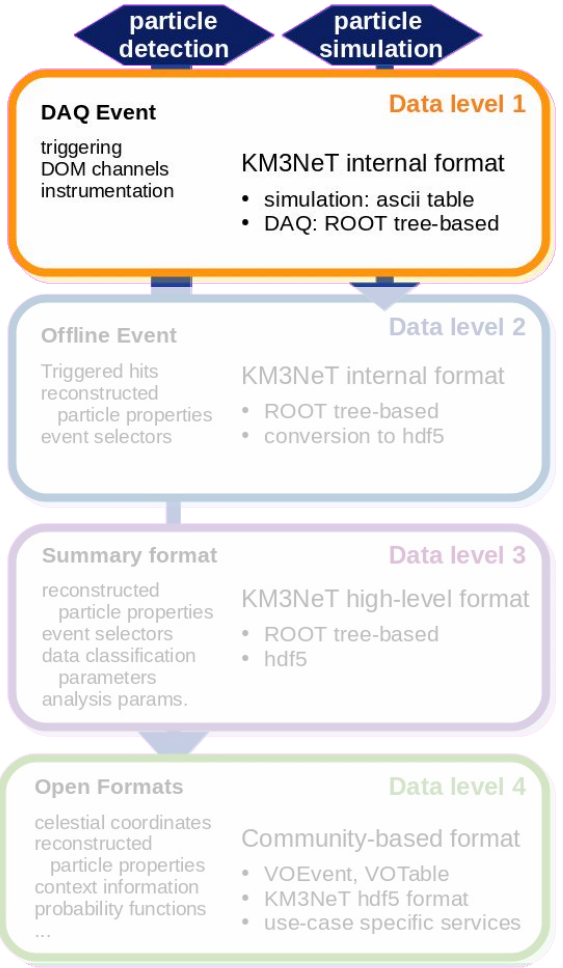
Primary product:
reconstructed (neutrino) event samples



Data processing pipeline (Tier 0)

Provenance on ***data generation***: state & configuration

- Internal Oracle Database to store information for each run file
  - Run configuration
  - Detector calibration
  - Trigger parameters
  - Additional meta parameters (e.g. run schedule mode and time, priority…)
- In-file provenance
  - ROOT UUID (each ROOT file contains a unique identifier)
  - Data writer node
    - Host
    - DAQ software version
    - ROOT version
    - UTC time



Data processing pipeline (Tier 1)

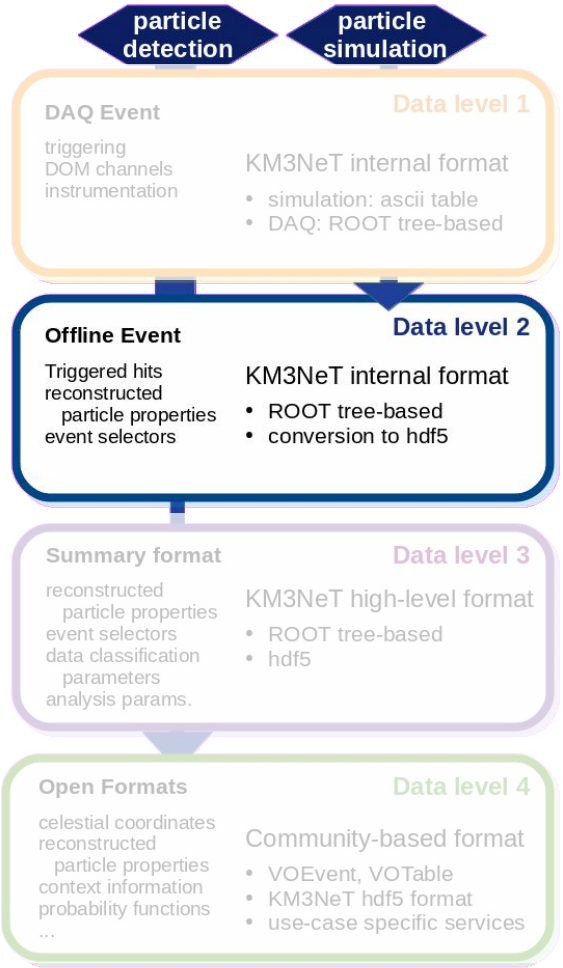
Provenance on ***data processing***: processing step, configuration, environment

Tier 1: reconstruction level including MC processing chains (ROOT-based)

- In-file provenance
  - ROOT UUID
  - Processing node
  - Application name
  - Command line options
  - Additional information about the host and software environment

KEY: TNamed JPrefit;1 GIT=12.1.0
ROOT=5.34/38
application=JPrefit
command=/pbs/throng/km3net/src/Jpp/v12.1.0/out//Linux/bin//JPrefit -a /sp….
namespace=KM3NET
system=Linux ccwsge1332 3.10.0-1062.9.1.el7.x86_64 #1 SMP Fri Dec 6 15:49:49 UTC 2019 x86_64

provenance information of 1 processing step



High-level analysis (Tier 2)

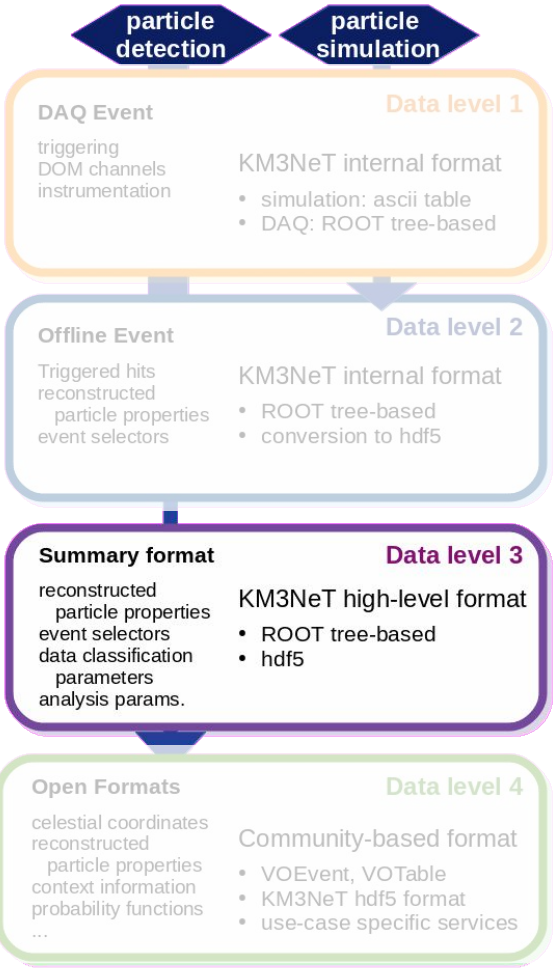
Provenance on ***data processing***: processing step, configuration, environment

Tier 2: event & parameter selection, conversion to open format (python-based)

- implemented in km3pipe
- tracking of pipeline steps
- preserving the full Python environment (incl. exact package versions)
- based on thepipe with extended provenance functionality
- storage as JSON in output file


```
{
  "uuid": "70980622-a936-4723-a63c-a83b3a128678",
  "name": "pipeline",
  "parent_activity": "a06e1bba-39f7-41f4-9228-eced9af611a6",
  "child_activities": [],
  "start": {
    "time_utc": "2020-08-31T11:15:00.114387+00:00",
    "peak_memory": 246.23046875
  },
  "stop": {
    "time_utc": "2020-08-31T11:15:00.415748+00:00",
    "peak_memory": 249.859375
  },
  "system": {
    "thepipe_version": "1.3.2",
    "executable": "/usr/bin/python3",
    "arguments": [
    "environment": {
    "platform": {
    "python": {
    "start_time_utc": "2020-08-31T11:15:00.144660+00:00"
  },
  "input": [
    {
      "url": "datav5.40.jorcarec.aanet.00006151.root",
      "comment": "OfflinePump input"
    }
  ],
  "output": [
  "samples": [],
  "status": "completed",
  "configuration": {
    "planned_cycles": 1,
    "modules": [
      {
        "name": "OfflinePump",
        "parameters": {
          "filename": "datav5.40.jorcarec.aanet.00006151.root"
        }
```


example output of tier 2 pipeline

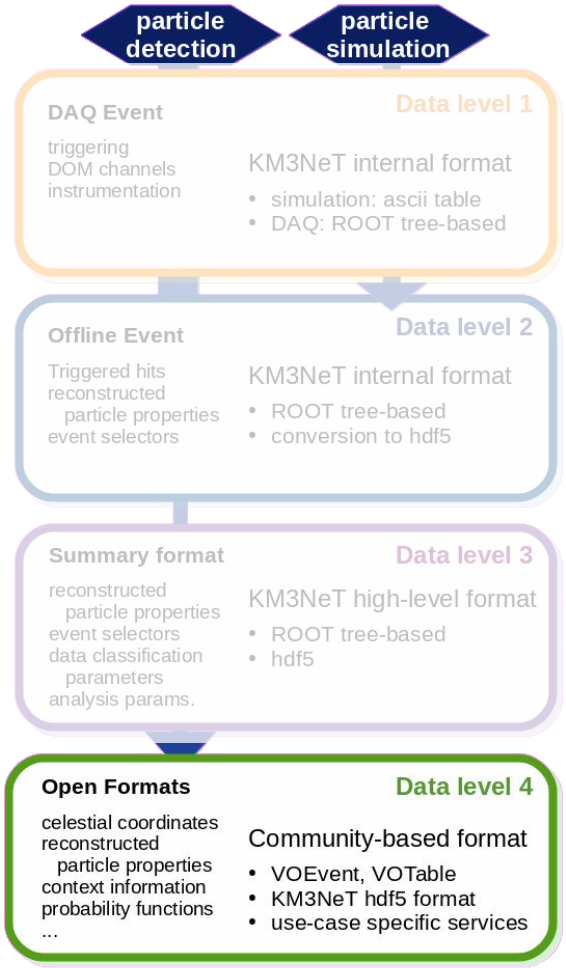


Provenance in open data format

Reduced provenance information for interested user

- for level 4 event files, (reduced) provenance information is stored in header information as hidden attributes
  - on data generation and conditions
  - processing steps
  - event and parameter selection
- full provenance scheme should be accessible through UUID relations and centrally stored as full provenance information

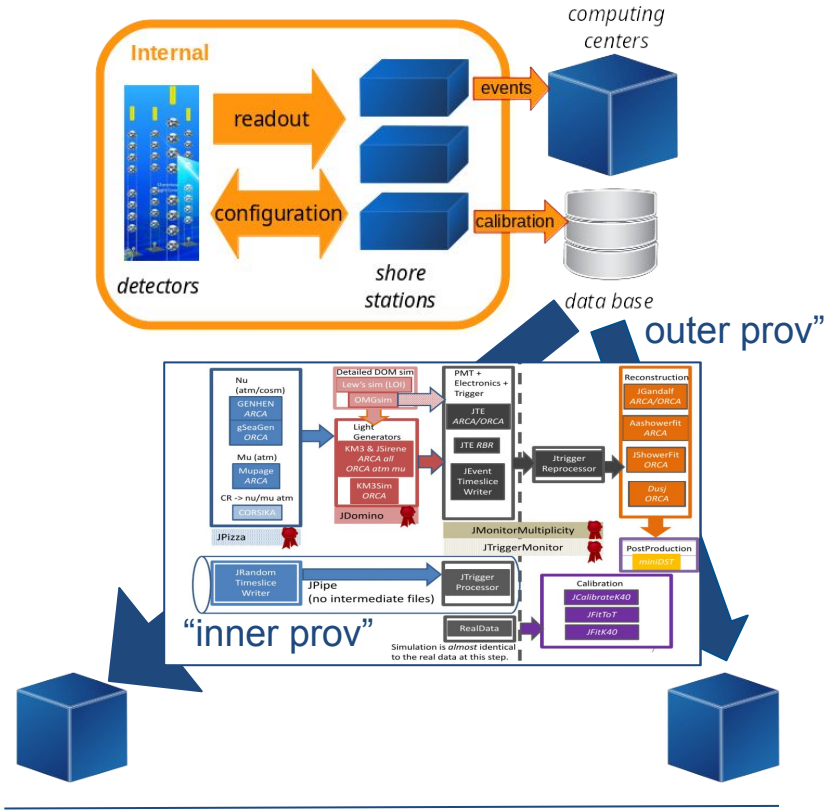
→ **database system** for full tracking of the processing chain and related information to the final data set



Requirements for full processing chain

- distributed computing workflow with flexibility of porting to different computing centres, especially in the Tier-2 regime
- Description and versioning of all input, output and configuration to processing steps
- tracking of inputs, outputs and full processing configuration

→ Not only “outer” black-box provenance but also “inner” gray-box provenance needed



Exploring CWL and Dirac

- Common Workflow Language could serve to describe the workflow (inner prov)
  - full annotation of the workflow possible
  - cwlprov as module to capture full “inner” provenance, using containerized software
- File processing backend in DIRAC interware, which does not provide full provenance per se (extension needed), but provides database for “outer” provenance
- Current questions to answer:
  - Semantic mapping might be an issue (especially for dynamic, data-driven HTC workflows)
  - Automatisation is a main goal, which might contradict CWL as it seems designed for static workflows?

COMMON WORKFLOW LANGUAGE

+

DIRAC THE INTERWARE

=

PROV ?

Summary

- First steps taken to storing of provenance information at all levels of data processing
- Needs formalization and streamlining of formats and information
- Currently in-file provenance, not meeting requirements for large-scale decentralized computing
- Exploring capturing of workflow description (CWL?) and middleware for distributed computing (DIRAC)
- Working towards full provenance scheme, user to be provided reduced information with possibility to retrieve full provenance information from database (for internal users)

Thank you for your attention!

ERLANGEN CENTRE
FOR ASTROPARTICLE
PHYSICS

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG