Evolution of the datasheet concept

E.Lancon
M.Potekhin

DAP Round Table
August 13 2026

Overview

- The project ”Preparing QCD Data for Foundation Models” includes three thrusts:
  - Detector-level data
  - LQCD
  - DAP
- Our current collaborative effort with JLab focuses on the DAP item. As a result of this collaboration, there is now a promising datasheet template for describing analysis-level data from a variety of experiments. We at BNL are using a select dataset from STAR as a test case.
  - Currently YAML is being used as the formatting medium
- Recently, after a dialogue with our colleagues in STAR and PHENIX, we reached a new understanding (of the modality) of the datasheet use, which we’ll cover in the following slides.

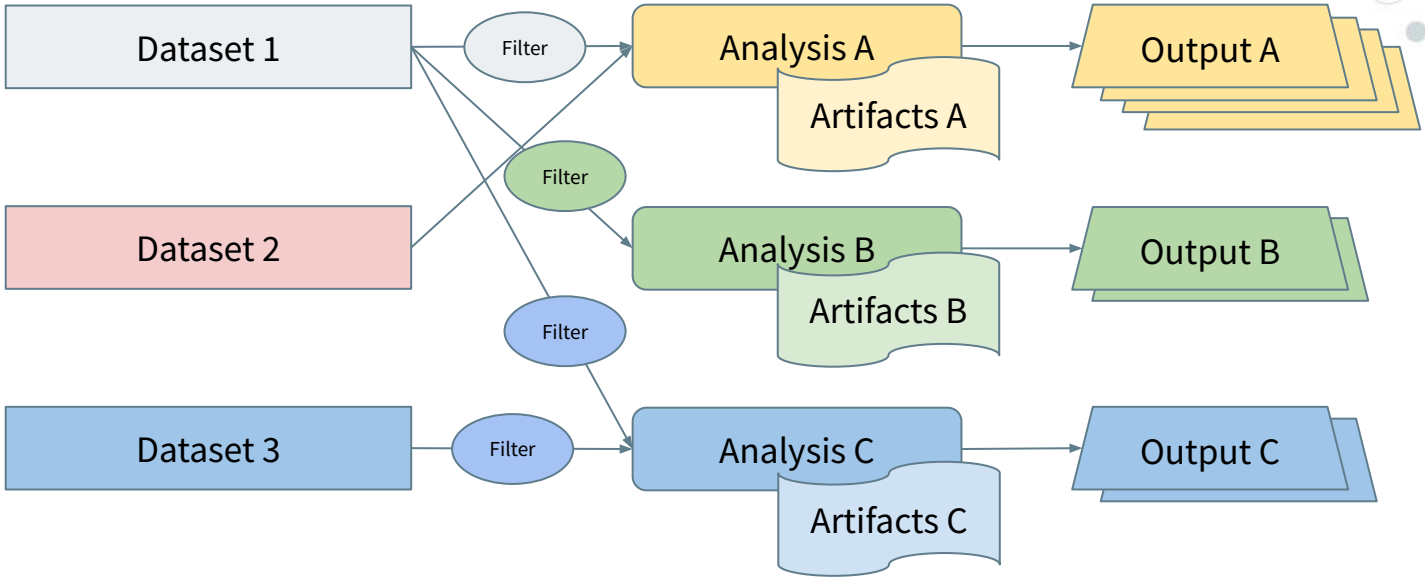
An old adage from the DAP Community…

- "Data is useless without software…"
- One can also say that "software is useless without provenance, versioning, configuration, workflow and data artifacts".
  - …often same applies to functional infrastructure e.g. databases
- There elements together form the analysis chain. The knowledge of how it's done and what are the dependencies is central to DAP.

What is a Datasheet?

- **Wikipedia:** “A datasheet, data sheet, or spec sheet is a document that summarizes the performance and other characteristics of a product, machine, component (e.g., an electronic component), material, subsystem (e.g., a power supply), or software in sufficient detail that allows a buyer to understand what the product is and a design engineer to understand the role of the component in the overall system.”
- In the first approximation, a datasheet is a collection of metadata to make it discoverable and reusable… But what about its “role in the overall system”?
- The “overall system” in our domain is the complete analysis chain. It has its own complex set of characteristics which in principle extraneous to the dataset as it is.

Datasets and analyses – it's not a one-to-one mapping



Analyses have different physics goals, workflows, event selection etc.

Enter the graph…

- The diagram on the previous slide is obviously oversimplified but…
- It is an illustration that in the DAP context, all these entities are related and these relations itself are crucial for the data to be (re)useable.
- Recording all this information in a datasheet describing a dataset is extremely cumbersome and not practical.
- Example: STAR is using a selection process to fetch sections of a given (large) dataset that conform to the needs of a specific analysis. Each analysis can (and often does) has its own “good run list” and/or “dead channel map”, both of which are analysis-dependent. These artifacts to not organically fit into a relatively flat structure of the datasheet.
- Proposed solution: model research process as a graph of **Research Objects**.

Analysis as a research object

- Analysis is a transformation of data, not a part of a metadata of some dataset.
- It has its own lifecycle, versioning, and a record with a citeable reference (DOI)
- It is linked to other research objects:
  - Datasets
  - Data artifacts specific to the analysis
  - Workflows
  - Documentation (knowledge preservation)
  - Software
  - Calibrations
  - Derived datasets
  - Publications
  - …

This research object has:

- Its own record
- Its own DOI
- Its own version control

Dataset vs Analysis research objects: the record metadata

Dataset **D1**

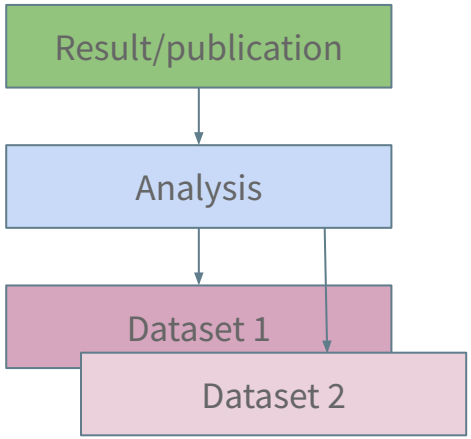
- Core conditions
  - Collision system
  - Energy
  - Detector configuration and condition
  - Magnetic field
  - etc…
- Location, size etc
- Citation
- Relationships
  - used_by_analyses:
    - A1
    - A2

Analysis **A1**

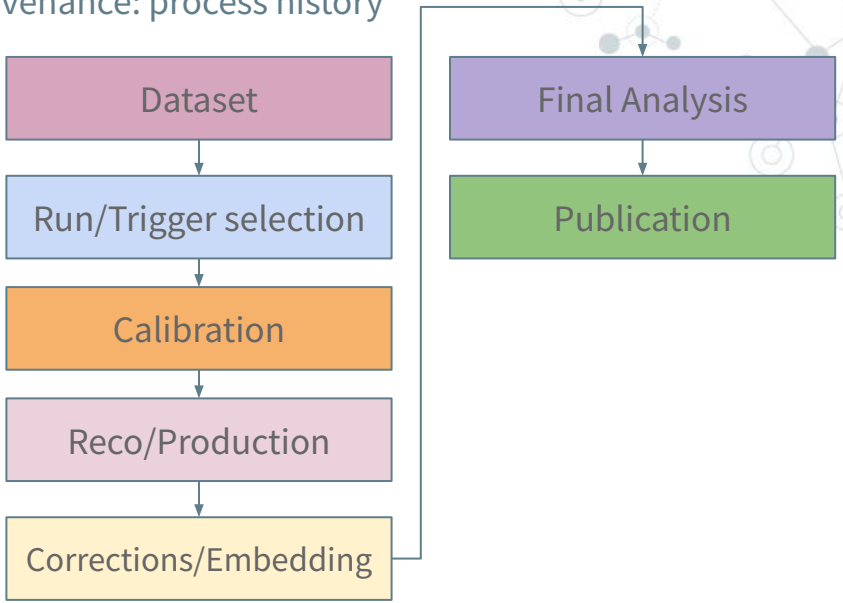
- Core metadata
  - Identity
  - Configuration and version
- Relationships
  - uses_datasets:
    - D1
  - Uses_workflow: W1
  - uses_software
    - S1
    - S2
  - Produces…
  - Publications:
    - P1

Citation vs provenance

Citation: declaration of dependencies



Provenance: process history



Core idea: the Research Object is a common abstraction

- This applies to
  - Datasets
  - Analyses
  - Data artifacts and conditions
  - Software items
- Every research object has
  - Identity: a DOI
  - Metadata: describing what it is, possibly with indexing
  - Lifecycle: version control over time.
  - Relationships: links to other objects (e.g. via DOIs)
- Scalable, reusable, FAIR-oriented and AI-friendly
- A natural way to manage the complexity of the research process
- Research is a graph of research object