Analysis Traceability and Provenance - DPHEP

Dr J Shamdasani, R McClatchey, A Branson and Z Kovacs

Contact : jet@cern.ch

University of the West of England

Outline

- Provenance
- CRISTAL
- Analysis Provenance and Neuroscience
- Provenance in N4U
- Applications for HEP

Provenance

- A Computer Science concept *(Wine, Meat, Art)*
- “Source or origin of a piece of data”
- It is a trace of how a “thing” or “entity” came into being
- It is an *audit trail* of how data came into existence *(benefit?)*
- W7 : Who, What, When, Where, Which, Why, How

Provenance

who ran an analysis, this is a user name,

for **what** purpose, what their analysis is supposed to achieve,

when they ran it this is a timestamp which denotes when it started and when it finished,

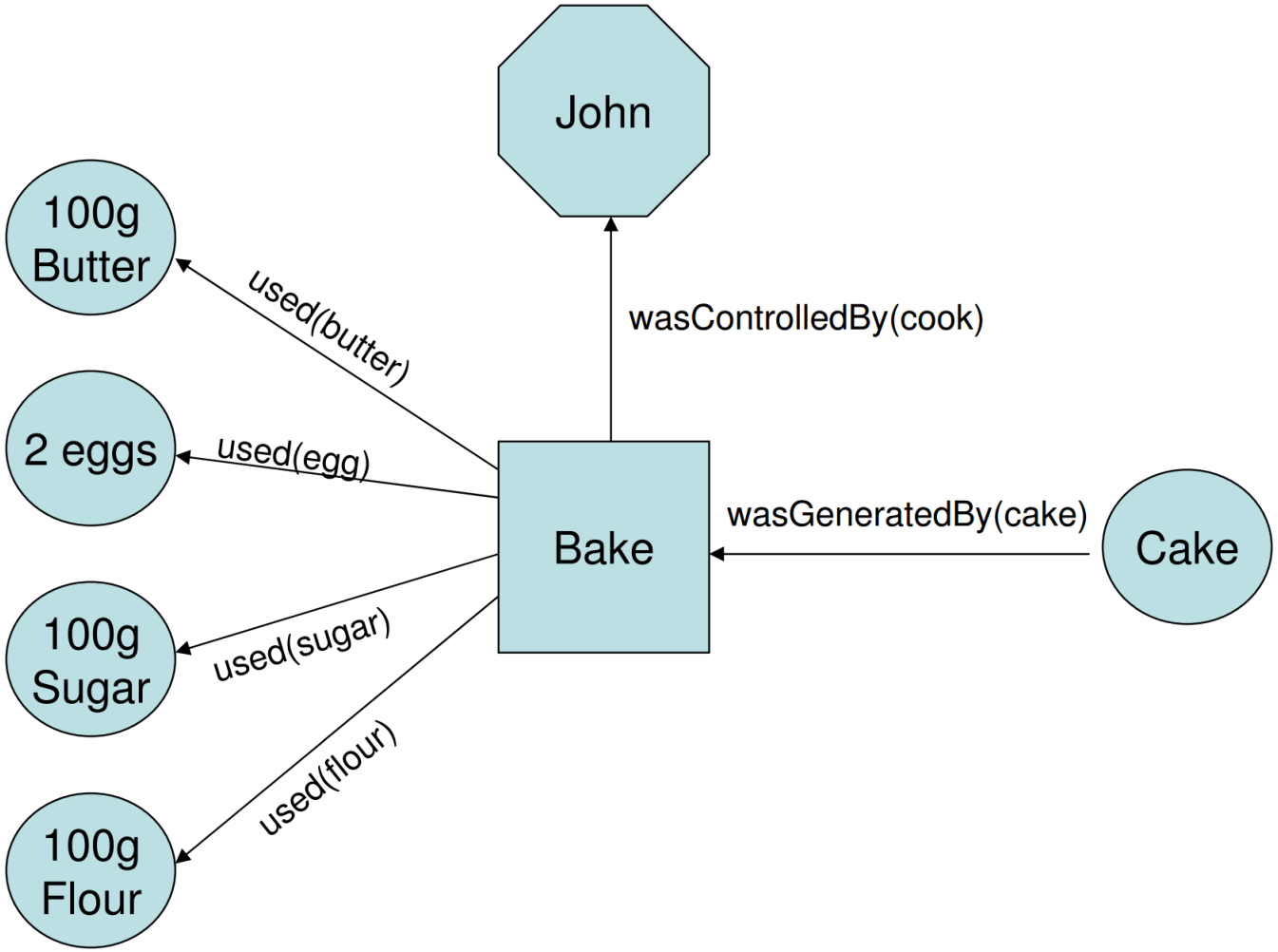
where it was run this is GRID and Cloud related information,

which datasets and algorithms were used to create and run their analyses,

how it was executed, this more detailed infrastructure information

and lastly **why** the analysis was run, this is a justification from the user.

Provenance : Example



CRISTAL

- Developed at CERN in early 2000s
- Used for the tracking of the CMS ECAL Detector
- A long history and pedigree
- Is provenance enabled by design
- Used in industry (BPM, Data Processing, R&D prototyping and production)

CRISTAL

- Takes a *meta-schema* approach
- This means that objects are *described* instead of instantiated
- These descriptions are stored as data in the system
- They are versioned and can be accessed at any time
- They can also be forked

Construction Provenance

- CRISTAL was created to track the construction of the CMS ECAL Detector
- The characteristics and identity of the components of the ECAL were gathered as structured, queryable data
- This provided quality control, decision support and eventually data for detector calibration

Analysis Provenance

- CRISTAL for computational research
- Developed for neuroimage analysis for the NeuGRID EC FP7 project and its follow-on N4U
- Used to track the production and the running of analyses on the GRID

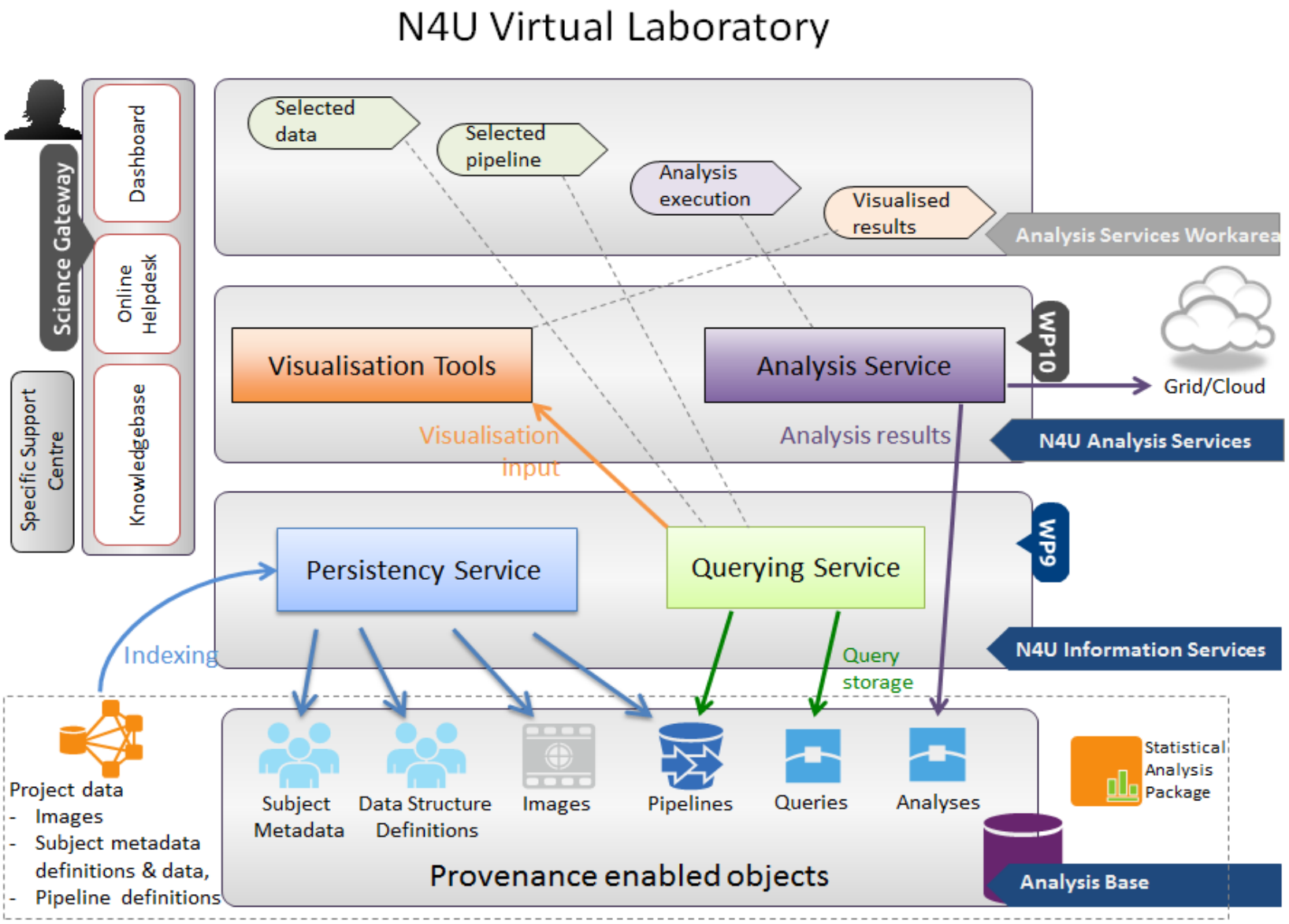
Neuroscience

- Analyses as workflows
- Therefore it is *workflow provenance*
- Events generated at step execution
  - These generate *metadata* which can be queried
- Provenance collected at *infrastructure* level as well

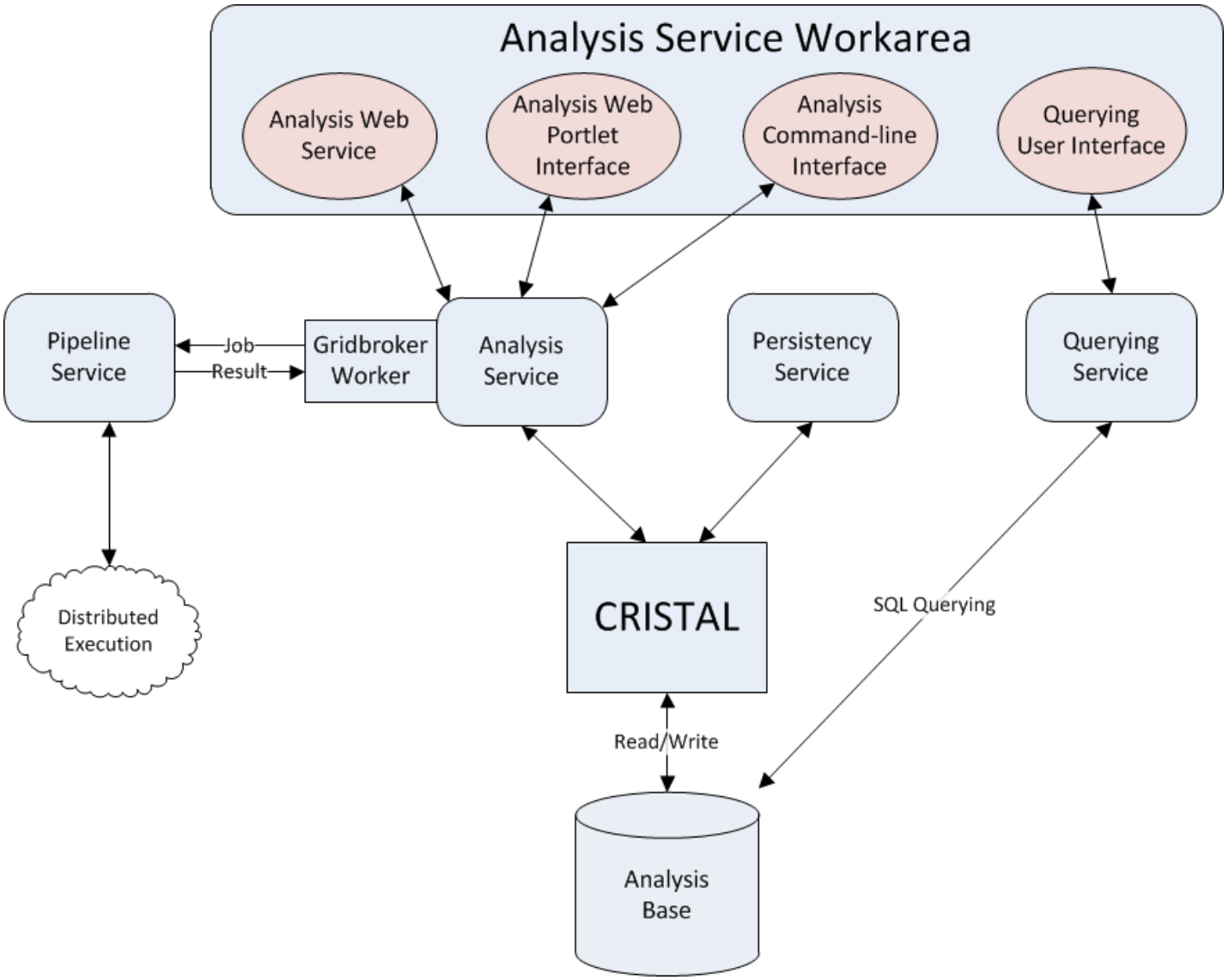
N4U

- Neuroscientists run 1000+ experiments a year
- They need to share results
- Provenance is key for this
- Datasets registered :
  - Images catalogued using clinical metadata
  - Usage tracked

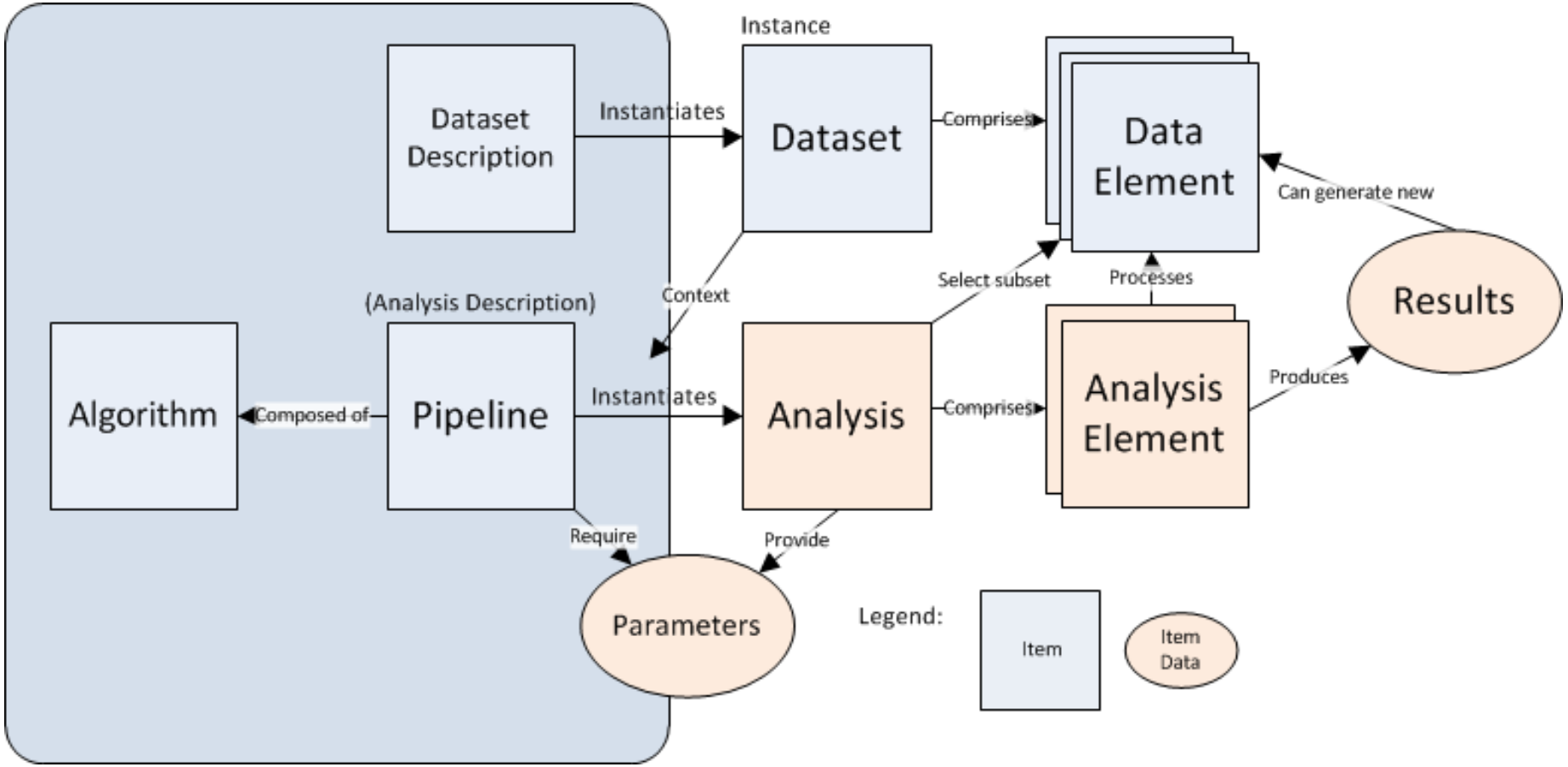
Provenance in N4U



Provenance in N4U



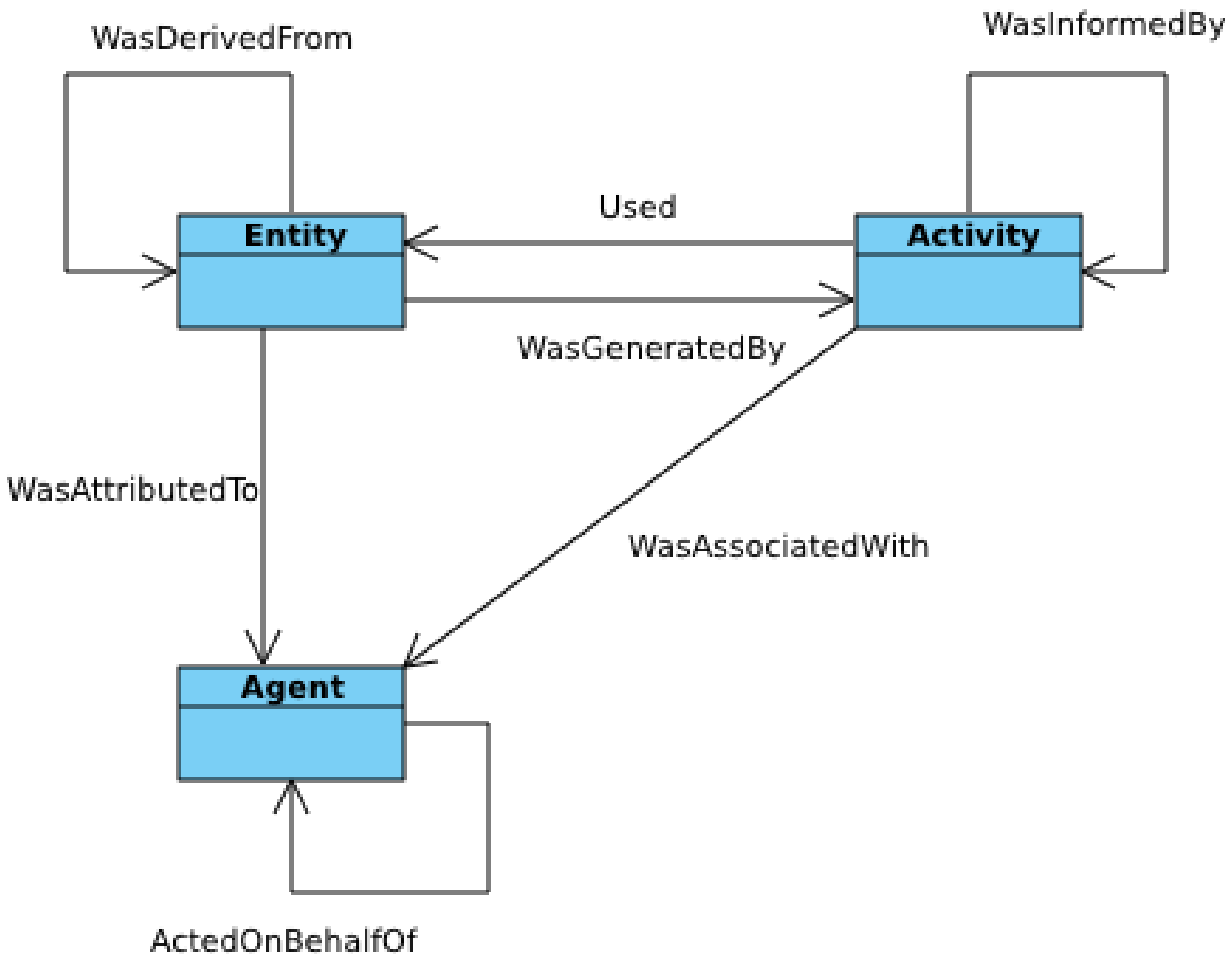
Provenance in N4U



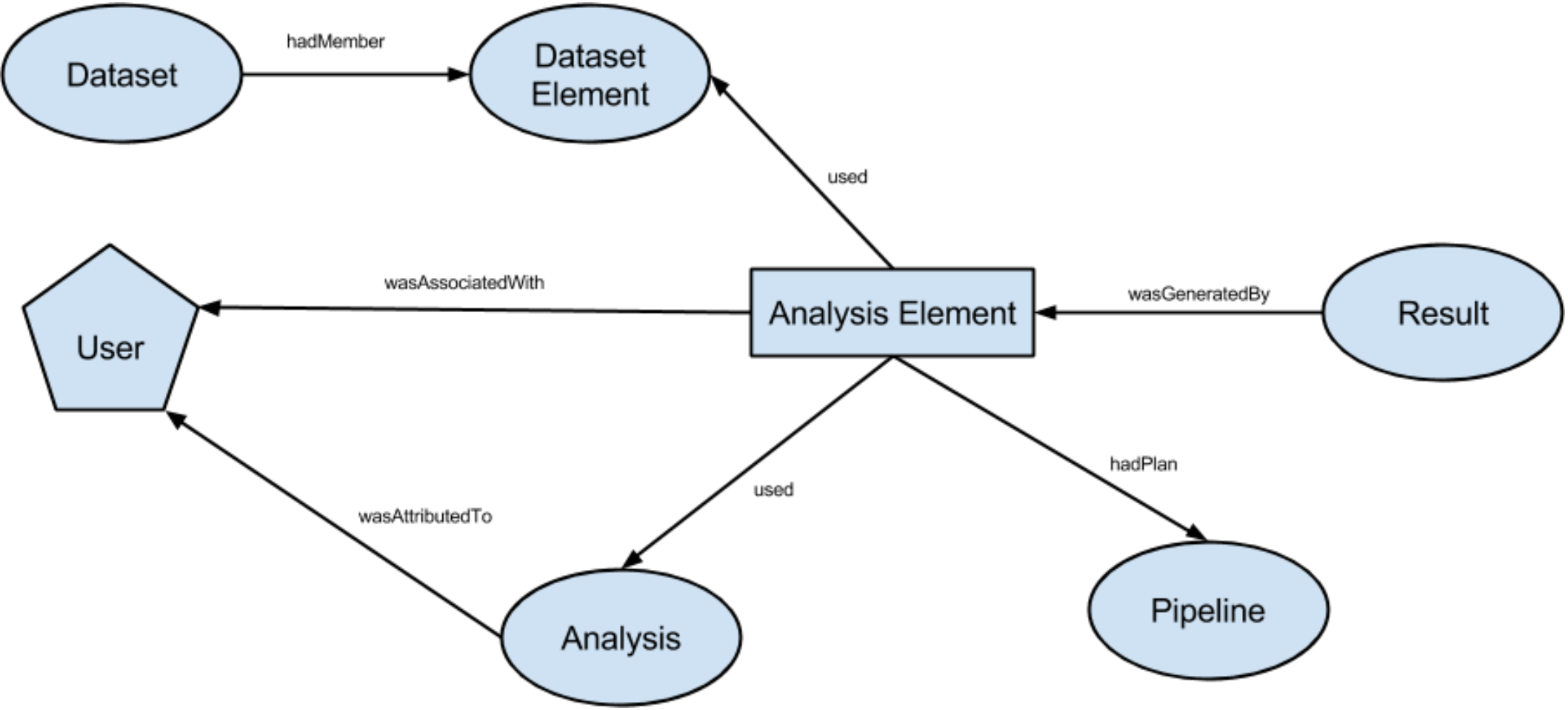
PROV

- W3C Standard for provenance *interoperability* can be used to capture provenance – not ideal!
- Three main top level classes : Entity, Activity, Agent
- Seven main relationships at the top level
- Can be serialized to different formats
- Its *extensible* – Which is key!

PROV – Top Level



Mapping to PROV



FOR HEP

- Currently working with the DPHEP initiative
- Applying “Provenance Enabled Objects” to the world of HEP : Analysis Provenance
- Future-proof dataset preservation though structure description and annotation.
- Work is currently ongoing

Conclusion

- CRISTAL is now open source – LGPL v3 (http://www.cristal-ise.org)
- Source Code : http://cristal-ise.github.io
- Used in Industry :
  - Technoledge (Geneva, Switzerland)
  - M1i (Annecy, France)
  - Alpha-3i (Rumilly, France)
  - New UK startup for dataset tracking