ATLAS Data Analysis Ontology: ontological representation of investigations

DKB Meeting

PanDA Workshop Apr 21-22 2016 at CERN

- Goals of DKB:
  - This was in the context of a conversation on whether we can/should work to capture and present the whole process from physicist idea ➞ production intent ➞ production request ➞ production status ➞ completion of the full processing chain ➞ available data
    - With ability to drill down within that chain for processing status, data availability, configuration and so on, drawing on content in prodsys, AMI, Rucio etc
- Data Sources:
  - PanDA/JEDI, DEFT, Rucio, AMI, AGIS, JIRA, googledocs, ...
- Currently R&D project, proceed to realistic use cases after a May technical/planning meeting at CERN

CISP: Core Information about Scientific Papers

- Identifies the main components of scientific investigations
- Metadata about the content of scientific papers

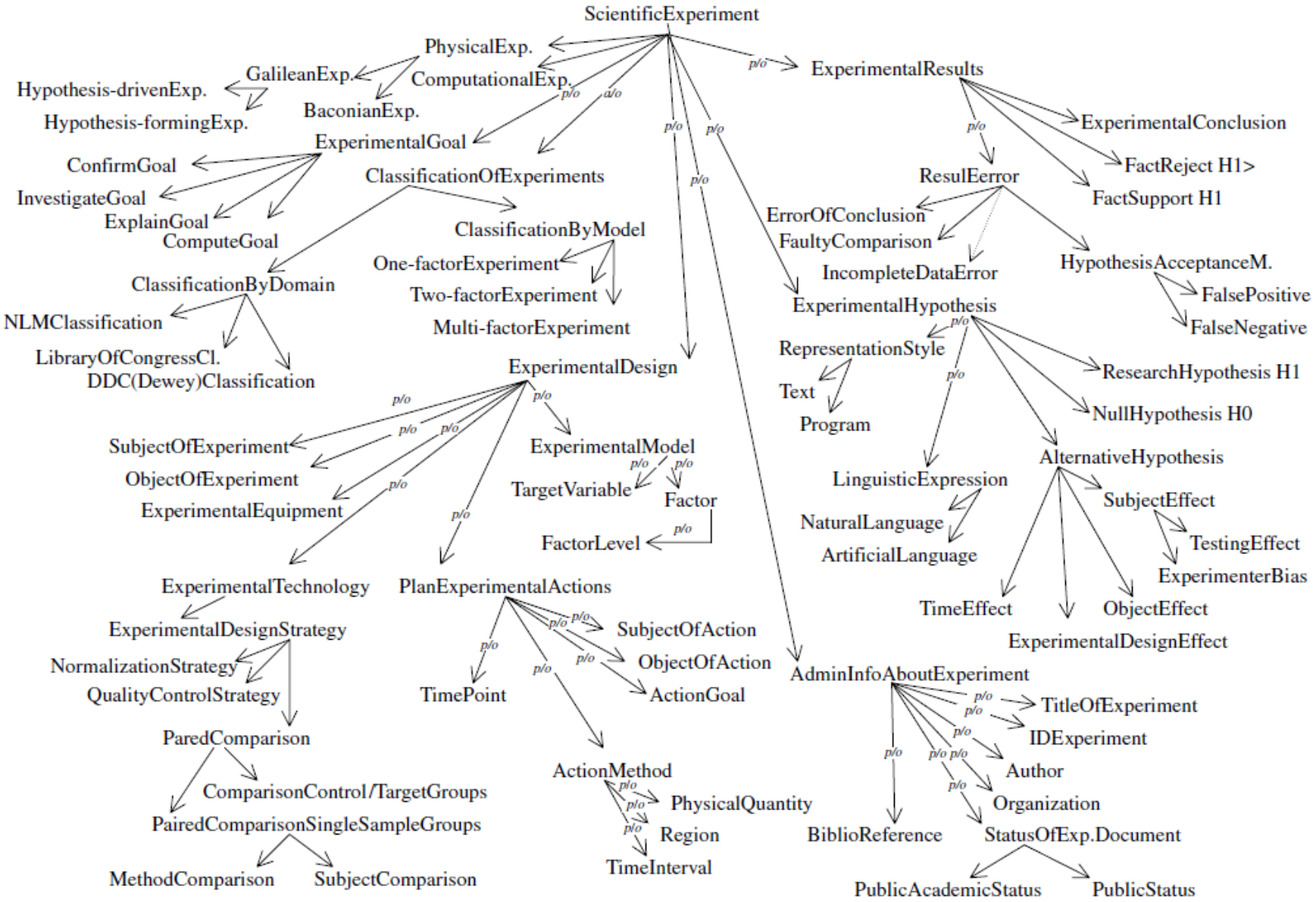
- CISP classes:
  - Goal of Investigation
  - Motivation
  - Object of investigation
  - Research Method
  - Experiment
  - Observation
  - Result
  - Conclusion

EXPO: Ontology of scientific experiments

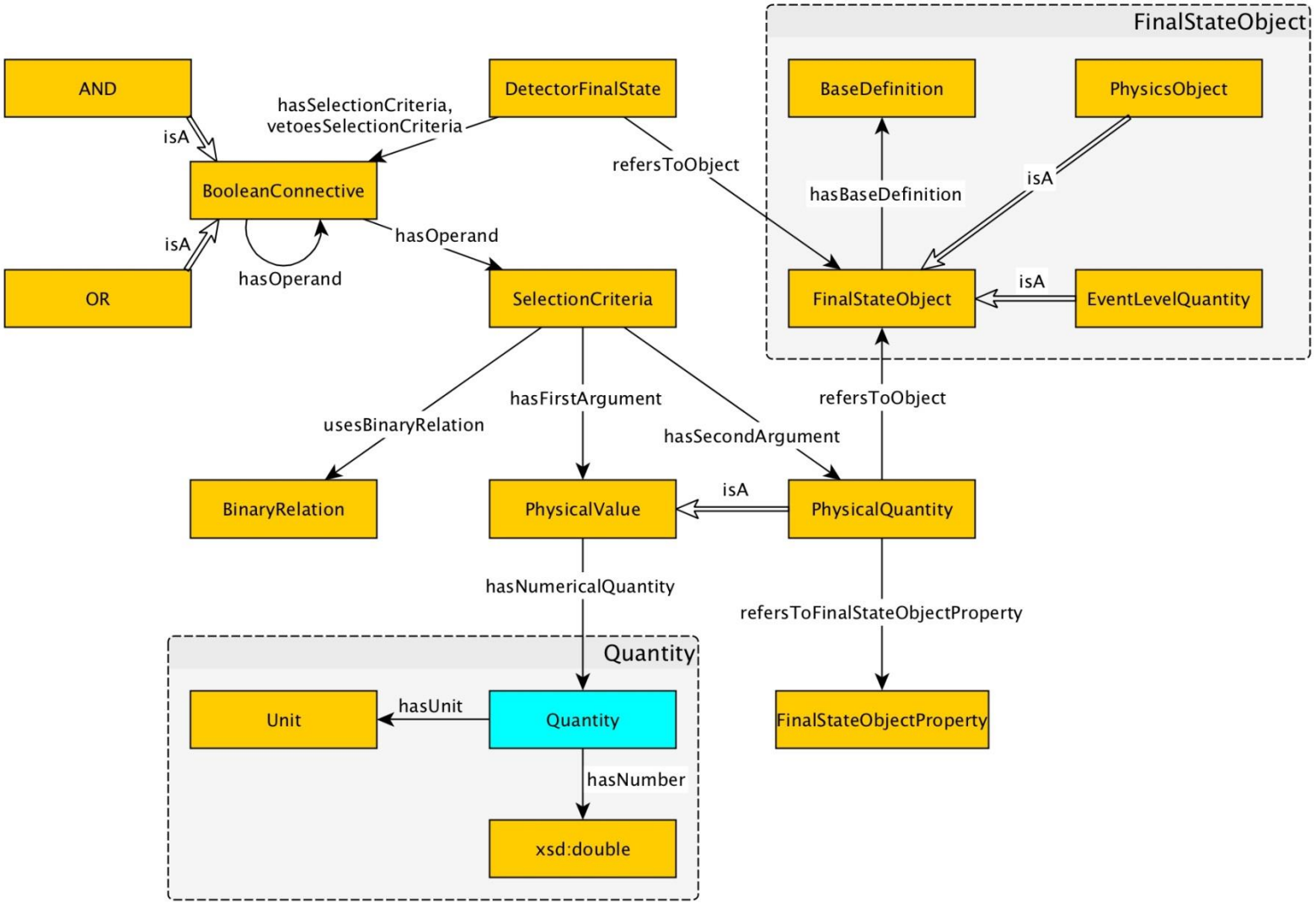
- A very complete ontology about scientific experiments proposed proposed at the University of Wales, Aberystwyth
- ~200 concepts to allow providing a formal description of experiments for efficient analysis

- EXPO classes:
  - Experimental design strategy
  - Fact
  - Field of study
  - Procedure
  - Variable
  - Experimental technology
  - Scientific activity
  - Hypotesis forming
  - Interpreting results

EXPO fragment



Detector Final State



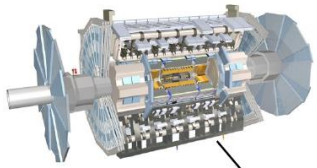
ATLAS Data Analysis Ontology Classes

- Data Analysis Description:
  - Data Analysis Unique Identifier [Ref.Code from GLANCE "TOPQ-2016-03"]
  - Short Title ["MS dilepton top mass at 8 TeV"]
  - Full Title ["Measurement of the Top Quark Mass in the dilepton channel from 8 TeV ATLAS Data"]
  - Analysis Team
  - Status (active, finished,…)
- Experimental model:
  - Detector Final State:
    - Luminosity [8 TeV, …]
    - Energy [20 femtobarn^{-1}, …]
    - Data Used [Run 1, Run 2, …]
    - Year [YYYY]
    - Other parameters
  - SW Release
  - ATLAS Data Samples
    - Monte Carlo data samples
    - LHC Data Samples
  - Data Processing Chain
    - LHC Data Processing
    - Monte-Carlo Data Processing
- Data analysis is classified by the following categories:
  - Physical topics (Standard Model physics, Higgs physics, SUSY, B-physics and light states, Exotics, Top physics, …)
  - Physical groups (FTAG, BPHY, REPR, TOPQ, ANALYSIS, Exotics, etc)
  - Sub Groups
  - Project (mc12_8TeV, mc12_13TeV, data10_7TeV, data11_2p76TeV, …)
  - Campaign/subcampaign
  - Domain of experiment (NML Classification, Library of Congress, DDC (Dewey) Classification)
- Experimental Results
  - Publications & Conf Notes

Experimental model

- Detector Final State:
  - Luminosity
  - Energy
  - Data Used
- SW Release
- Data Processing Chain
  - LHC Data Processing
    - Trigger
    - Reconstruction
    - Derivation
  - Monte-Carlo Data Processing
    - Generation
    - Simulation
    - Digitization
    - Reconstruction
    - Derivation

We can use DetectorFinalState ontology, proposed in this paper:
http://dase.cs.wright.edu/sites/default/files/WOP2015_pattern_abstract_5.pdf



ATLAS Data Samples

Data Samples attributes:

- Dataset ID
- Dataset type
- Project
- Stream
- Generator
- Data format
- Production step
- Number of files
- Number of events
- Total size

- Data samples
  - Signal
    - Production (ggF, VBF, VH, ttH)
    - Decay
  - Background
    - Category (Top, WW, ZV, ZX, Higgs)
    - Process ($t\bar{t}$+ boson, t + boson, WW, ZZ, …)
    - Generator (MADGRAPH, POWHEG, SHERPA, Pythia, Pythia8)

- Data formats
  - RAW
  - ADO
  - DADO
  - EVNT
  - HITS
- Dataset types
  - Real Data
  - Monte Carlo (MC)
- Triggers
  - The first level trigger (LVL1)
  - the second level trigger (LVL2)
  - the High Level Trigger (HLT)
- Streams
  - Muon
  - Egamma
  - …