Data Knowledge Base for HENP Scientific Collaborations

Maria Grigorieva, Marina Golosova, Torre Wenaus, Alexei Klimentov

HENP Metadata Issue

HENP experiments metadata sources exist independently. Some metadata are integrated, but in general scientists need to obtain cross relations among metadata within each category and among categories by themselves.

DKB Basic Consideration

Organizing metadata in ATLAS, so as to provide a holistic view on physics topics, including integrated representation of all ATLAS documents (papers, drafts, supporting documents, conference notes, meetings, collaborative pages, etc) and corresponding data samples.

DKB Project Evolution

Started in the May of 2016 year (kick-off meeting at CERN and the first R&D phase):

> *"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" [Torre's talk]*

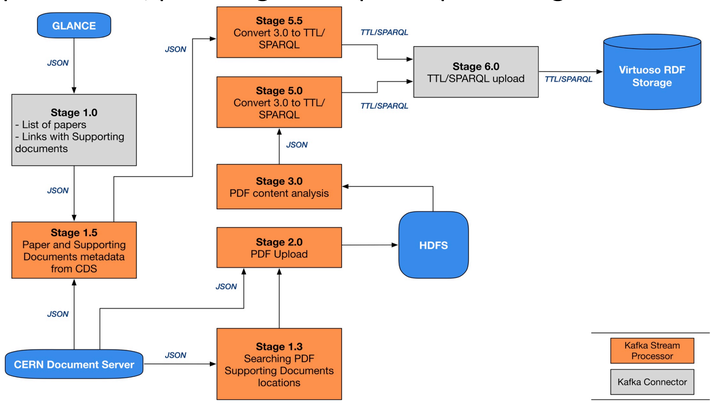
Implemented the first DKB Prototype, based on ontological metadata representation in Virtuoso RDF-Storage. That prototype provided links between ATLAS Publications and supporting documents with the corresponding data samples.

April – May 2017 discussions within ATLAS

- *DKB had been suggested to become a part of ATLAS Dataset Curation and Characterization Project (DCC) as a searchable system, providing flexible search for metadata and for metadata cross relations in various data sources.*
- DKB was accepted as ***ATLAS R&D Project.***

Apache Kafka Streams for Dataflow Automation

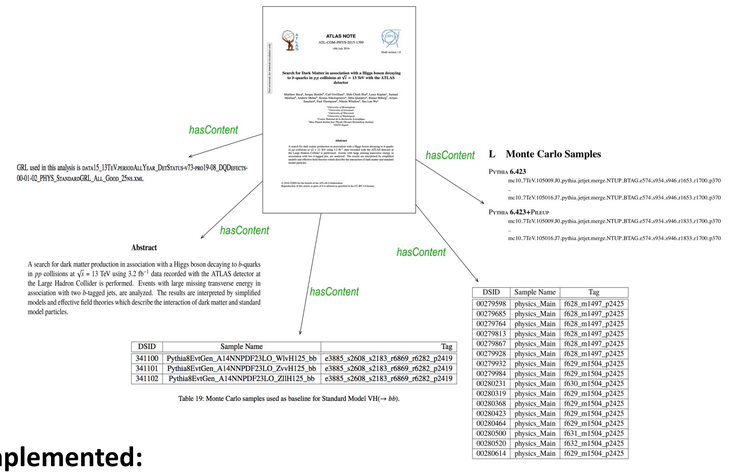
Specific services – metadata export/extraction/import tools, aggregation/integration modules – are organized as workflows run by Apache Kafka, providing nonstop data processing.



Implemented:

- Adapters to reuse existing data processing (extracting, transforming, uploading, …) modules within Kafka-driven dataflow
- Management utility to run the dataflow stages
- Formulated basic rules to write dataflow modules

Metadata extraction from text of ATLAS Internal Notes



DSID	Sample Name	Tag
341100	PythiaBEvtGen_A14NNPDF23LO_WlvH125_bb	e3885_s2608_s2183_r6869_r6282_p2419
341101	PythiaBEvtGen_A14NNPDF23LO_ZvvH125_bb	e3885_s2608_s2183_r6869_r6282_p2419
341102	PythiaBEvtGen_A14NNPDF23LO_ZllH125_bb	e3885_s2608_s2183_r6869_r6282_p2419

Table 19: Monte Carlo samples used as baseline for Standard Model VH($\rightarrow bb$).

DSID	Sample Name	Tag
00279598	physics_Main	f628_m1497_p2425
00279685	physics_Main	f628_m1497_p2425
00279764	physics_Main	f628_m1497_p2425
00279813	physics_Main	f628_m1497_p2425
00279867	physics_Main	f628_m1497_p2425
00279928	physics_Main	f628_m1497_p2425
00279932	physics_Main	f629_m1504_p2425
00279984	physics_Main	f629_m1504_p2425
00280231	physics_Main	f630_m1504_p2425
00280319	physics_Main	f629_m1504_p2425
00280368	physics_Main	f629_m1504_p2425
00280423	physics_Main	f629_m1504_p2425
00280464	physics_Main	f629_m1504_p2425
00280500	physics_Main	f631_m1504_p2425
00280520	physics_Main	f632_m1504_p2425
00280614	physics_Main	f629_m1504_p2425

Implemented:

PDFAnalyzer tool extracts metadata from TXT and XML representation of PDF text, by regular expressions and context analysis.

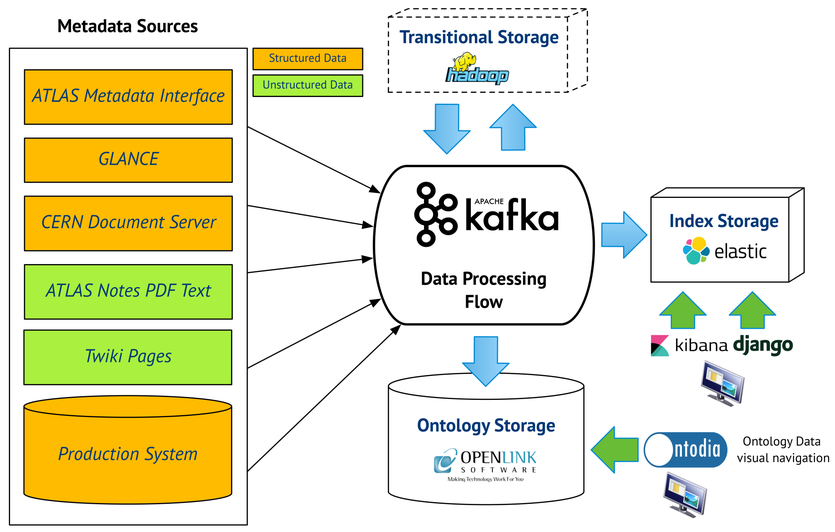
PDFAnalyzer extracts

- dataset names by regular expression
- datasets metadata from tables
- experiment-specific metadata from text

Returns structured metadata in ***JSON***

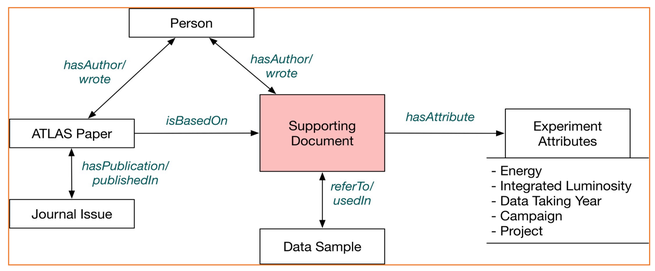
Has GUI interface, providing manual correction of analysis results

DKB Architecture Prototype

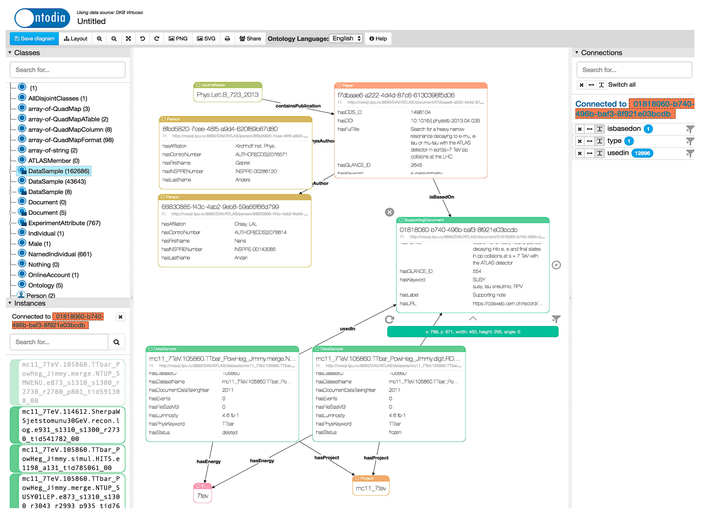


DKB program code on GitHub: ***https://github.com/PanDAWMS/dkb***

Ontological/Graph modelling

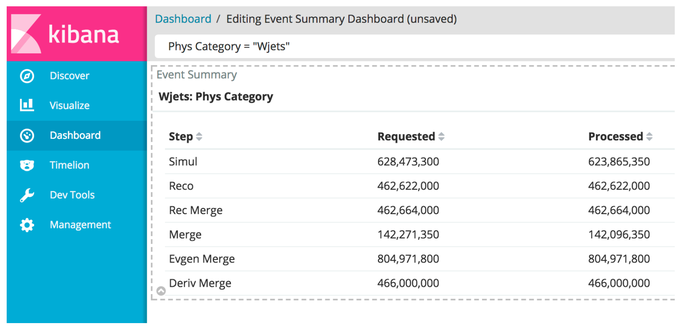


Ontology Storage Visual Navigation (Ontodia)



Indexing metainformation using ElasticSearch

- Metadata from Production System database are indexed in ElasticSearch, allowing to use hashtag's categorized aggregated event summaries



Reimplemented Twiki Event Summary Report, aggregated by Physics Category

Under Development:

(+) Fully automated metadata Kafka-driven data streaming

(+) DKB software transferring to CERN machines

(+) The development of the metadata crawler, extracting metainformation from Production System, BigPanDA monitor reports, Twiki pages, and put it in index storage

(+) User's request converter – implementation of google-like user request string

Search string: *'MC16a, Wjets, Powheg+Pythia8, w+ in mumu'*

(+) Datasets search by ATLAS geometry, conditions tags, Monte-Carlo campaign names, hashtags, physics channels and other parameters

(+) Metadata cross-relations investigation, using ElasticSearch and Kibana

(+) Graphical User Interface, based on Django

Metadata sources for DKB

1. ATLAS Management and Operation Framework

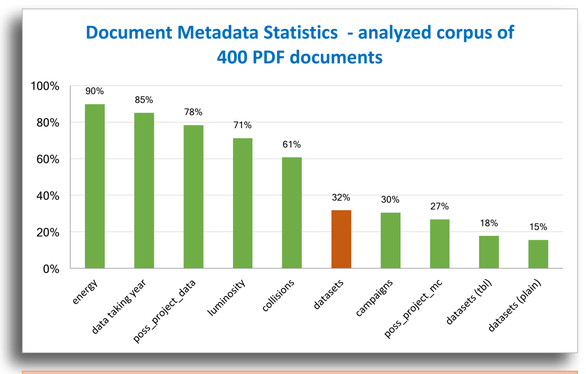
Allows to get a list of all ATLAS Publications with links to corresponding Supporting Documents

2. CERN Document Server

The main source of all Document's metadata

3. ATLAS Supporting Documents content

Unstructured PDF to structured JSON view representation



The information about datasets (in plain form or in tables), used in analysis was found only in 32% of analyzed documents

4. ATLAS Collaborative Documentation System (Twiki Pages)

Metainformation in Twiki pages provides the most suitable metadata categorization and representation for the end-users.

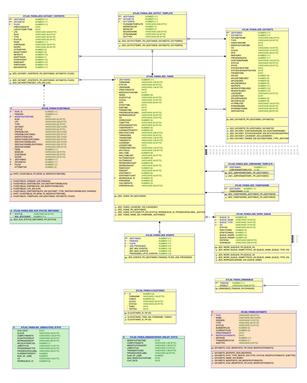
Twiki Monte-Carlo Campaign Pages contains:

- Aggregated events reports, categorized by physics categories
- Data sample's lists for each physics category
- Data sample's lists with breakdown a set of parameters, like:
  - Monte-Carlo Generators
    - Powheg+Pythia8 …
  - Physics channel
    - W+ in mumu
    - W- in taumu
    - Z/gamma* in tau tau …
  - Filtration methods
    - Without lepton filter
    - Two lepton filter …

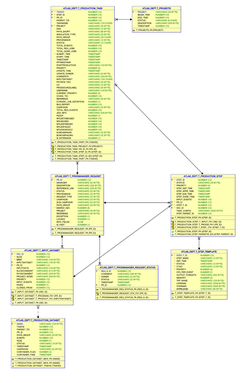
But the metainformation in Twiki is not enough structured and it doesn't provide mechanisms for synchronization with database back-ends.

5. Distributed Production Framework – ProdSys2

- **Request Interface:** allows production managers to define a request
- **DEfT:** translates user request into task definitions
- **JEDI:** generates the job definitions
- **PanDA:** executes the jobs in the distributed infrastructure



JEDI Database Schema



DEFT Database Schema

Beginning with the 2016th Monte-Carlo simulation campaign, ProdSys2 metadata were enhanced with '***hashtags***' for tasks, providing more detailed search by physics categories, physics channel, Monte-Carlo generators list, etc.

CP EG electron Epos EvtGen jet Lambda lepton MadGraphPythia mc MGPy muon PowhegPythia Pythia SameSign scalar tau valid VBFNLOPythia WarpedED

ACAT 2017 21-26 August, University of Washington, Seattle

BROOKHAVEN NATIONAL LABORATORY

UNIVERSITY OF RESOURCE-EFFICIENT TECHNOLOGIES — TOMSK POLYTECHNIC UNIVERSITY

National Research Center "Kurchatov Institute"