




Data Knowledge Base for HENP Experiments

Kurchatov Institute R&D Project

Progress Report

Marina Golosova

for NRC KI and TPU teams

Data Knowledge Base kick-off meeting highlights. May 2016

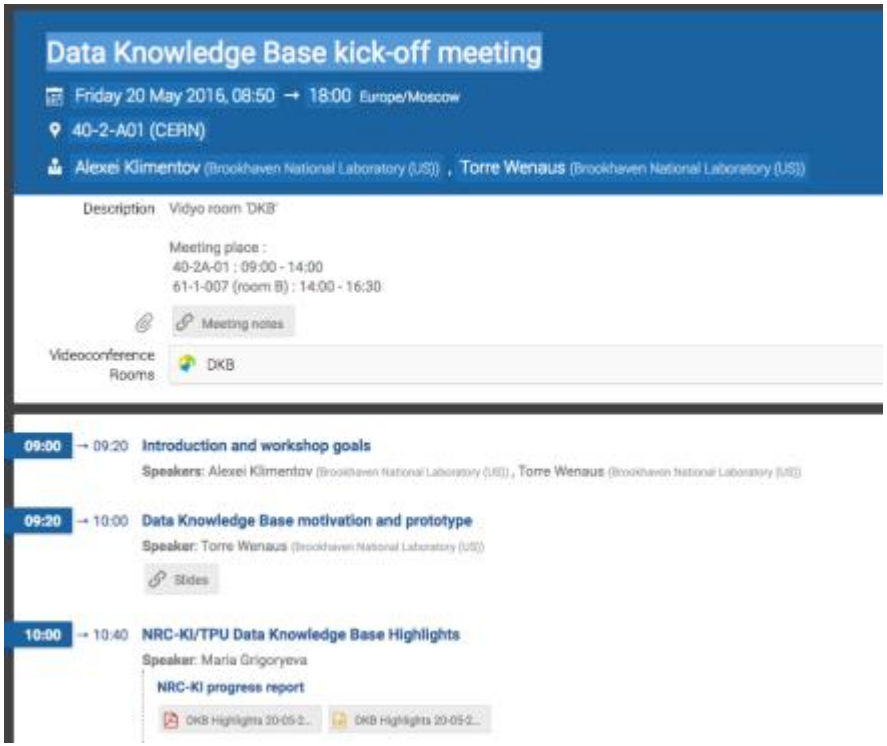
DKB Motivation

Torre Wenaus talk in May 2016

"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"

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, Indico meetings, Twiki pages, etc) and corresponding data samples (real data, MC datasets, containers).



https://indico.cern.ch/event/527581/

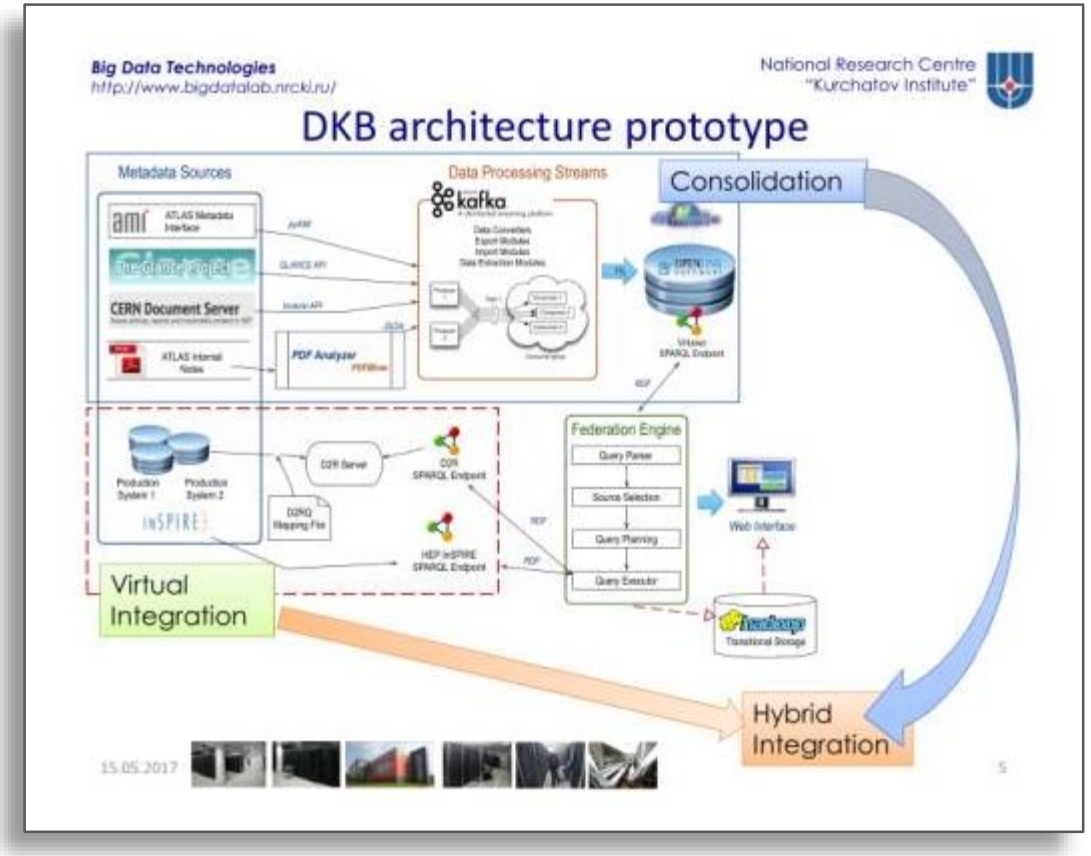
DKB is considered to look for cross references among the metadata, stored in various data sources.

slide by M.Grigorieva
for DKB/DCC meeting, 20.04.2017

DKB / DCC Technical Discussion. April 2017

- The first year of R&D was dedicated to the concept prototype development and technology evaluation
- PDFAnalyzer: unstructured documents as metadata source
- The prototype was presented at CERN a month ago (April 2017)
- Discussion conclusion: DKB project can go on as a part of DCC project, commonly known as DCC Whiteboard

Prototype conceptual architecture



https://indico.cern.ch/event/632634/

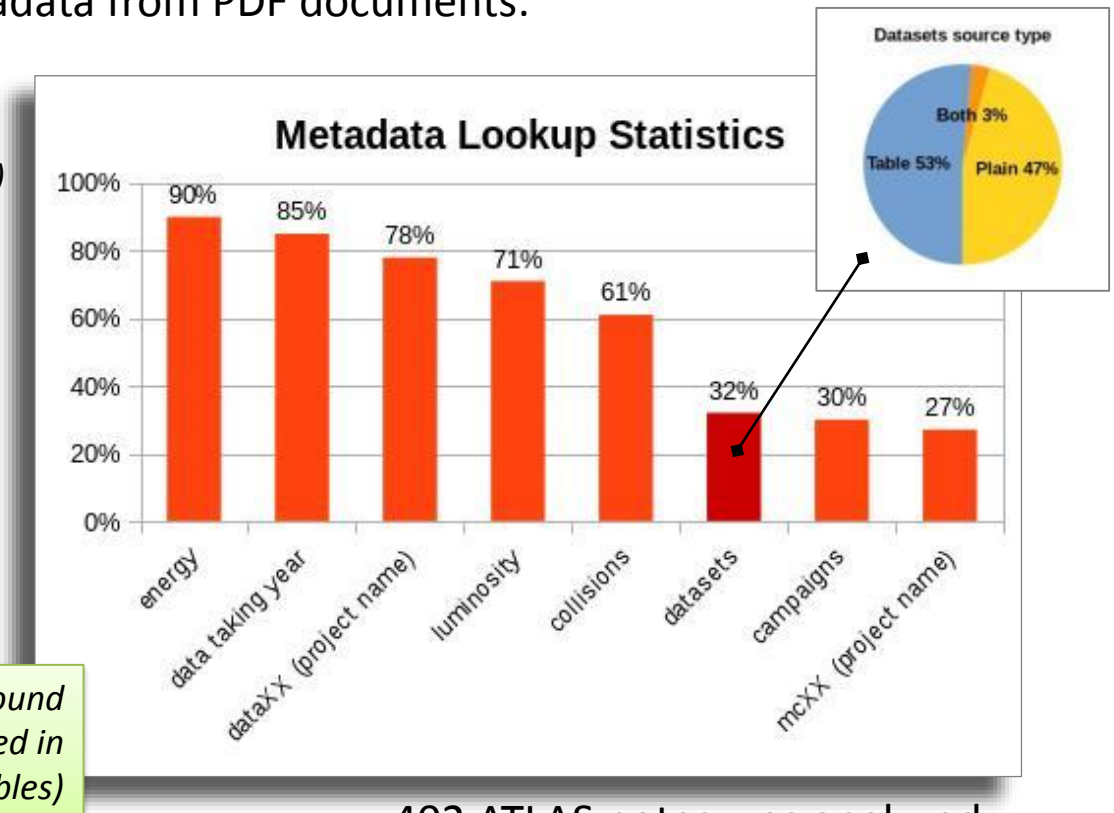
R&D project progress

- Infrastructure
  - DKB code moved from NRC KI Subversion to GitHub (https://github.com/PanDAWMS/dkb) *(NRC KI) (done)*
  - Deploy an instance on CERN resources *(…) (under investigation)*
  - SPARQL endpoint for ProdSys 1 & 2 DB *(TPU) (work in progress)* (to make ProdSys metadata available as a part of Semantic Web)
- Metadata Analysis
  - Dataset metadata overall view *(NRC KI) (work in progress)*
    - AMI/ProdSys/Rucio/Documents -> gathering metadata model, Semantic Web
    - Crosscheck
      - AMI/GLANCE metadata & papers (PDF) text analysis findings
  - Document content analysis – statistics *(NRC KI) (done)*

Documents content analysis

PDFAnalyzer extracts following metadata from PDF documents:

- Energy
- Data taking year
- RealData project name *(suggestion)*
- Integrated luminosity
- Collision type
- Datasets
- Campaign(s)
- MC project name *(suggestion)*
- Datasets (in table format)
- Datasets (in plain text format)



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

402 ATLAS notes was analyzed

Project manpower & roles

- Maria Grigorieva
  - project leader
  - core developer
- Marina Golosova
  - leading developer
  - technical support & system administration
- Alexandr Alexeev
  - technical support & system administration
- Vasily Aulov
  - SW developer, PDFAnalyzer author
- Maxim Gubin
  - Semantic Web & ontology expertize
  - technology evaluation

Institutes:
- TPU
- NRC KI

~2.2 FTE

Supported by:
Russian Ministry of Science and Education
and
Russian Science Foundation for Basic research
grants

Current tasks assignments

- Data Processing modules for extracting/processing/converting metadata from GLANCE, CDS, ProdSys
  - **Marina Golosova**
  - **Maria Grigorieva**
- Kafka Streams for automated data processing
- Hadoop Transitional Storage administration
- GitHub repository administration
  - **Marina Golosova**
- ATLAS Internal Notes PDF Analyzer
  - **Vasily Aulov**
- Ontological model of data analysis in HEP
  - **Maria Grigorieva**
  - **Maxim Gubin**
- Web Interface
  - **Maria Grigorieva**
- SPARQL endpoints (for ProdSys 1&2, AMI, …)
  - **Maxim Gubin**
- TPU Virtuoso Server administration
  - **Alexander Alexeev**

Institutes: TPU, NRC KI

Summary and Plans

First year major accomplishments

- Technology evaluation
  - RDF storage: Virtuoso
  - Transitional storage: Hadoop
  - Metadata streaming&processing: Apache Kafka
  - Knowledge Base navigation (Web GUI): Ontodia
- System prototype architecture was designed and implemented
- PDFAnalyzer was designed, coded & implemented

Second PY plans*

- Development
  - Enhanced internal notes search (CDS)
  - Fully automated metadata streaming
  - Federated prototype
  - Extended Web-interface (search tools)
- Technical points
  - Scalability tests
  - DKB with user authentication
    - *CERN SSO (under investigation)*

** To be tuned after today`s meeting*

Possible development directions

- Semantic Web for dataset metadata sources

  Cover existing dataset metadata sources with a Semantic Web
  - Single access point
  - Ongoing cross-check
- Semantic Web with feedback

  Add possibility to not only ***read*** through the Semantic Web envelop, but also to ***write*** back to the underlying systems
  - Filling gaps

    After discovering some missed in other sources information, store it to DKB storage (as it is going now) ***AND*** propagate it to the other sources
    And later DKB storage can be omitted for propagated data
  - User defined content

    Having a unified interface for a number of ATLAS metadata storage systems, one might like to alter found data or add some more

Thanks

- This talk drew on presentations, discussions, comments, input from many. Thanks to all, including those I've missed
  - Kaushik De, Dmitry Golubkov, Alexei Klimentov, Mikhail Korotkov, Dimitry Krasnopevtsev, Eygene Ryabinkin, Anatoly Tuzovsky,...
  - Special thanks go to Torre Wenaus who initiated this work and for his ideas about Data Knowledge Base content design

This work was funded by

the Russian Ministry of Science and Education under contract #14.Z50.31.0024

the Russian Foundation for Basic research under contract #16-37-00246.






BACKUP SLIDES

Cross-check: link Dataset to Document

Document: STDM-2015-05

- found by both AMI/GLANCE and PDFAnalyzer
- found only by AMI/GLANCE

data15_13TeV.00267358.physics_MinBias.recon.ESD.r6849
data15_13TeV.00267359.physics_MinBias.recon.ESD.r6849
data15_13TeV.00267599.physics_MinBias.recon.ESD.r6849
mc15_13TeV.119995.Pythia8_A2MSTW2008LO_minbias_inelastic_low.recon.ESD.e3432_s2081_s2132_r6108
mc15_13TeV.119995.Pythia8_A2MSTW2008LO_minbias_inelastic_low.recon.ESD.e3432_s2081_s2132_r6109
mc15_13TeV.119995.Pythia8_A2MSTW2008LO_minbias_inelastic_low.recon.ESD.e3432_s2081_s2132_r6121
mc15_13TeV.119995.Pythia8_A2MSTW2008LO_minbias_inelastic_low.recon.ESD.e3432_s2081_s2132_r6124
mc15_13TeV.119995.Pythia8_A2MSTW2008LO_minbias_inelastic_low.recon.ESD.e3432_s2081_s2132_r6125
mc15_13TeV.361200.Pythia8_Monash_NNPDF23LO_ND_minbias.evgen.EVNT.e3639
mc15_13TeV.361200.Pythia8_Monash_NNPDF23LO_ND_minbias.recon.ESD.e3639_s2601_s2132_r6616
mc15_13TeV.361201.Pythia8_Monash_NNPDF23LO_SD_minbias.evgen.EVNT.e3639
mc15_13TeV.361201.Pythia8_Monash_NNPDF23LO_SD_minbias.recon.ESD.e3639_s2601_s2132_r6616
mc15_13TeV.361202.Pythia8_Monash_NNPDF23LO_DD_minbias.evgen.EVNT.e3639
mc15_13TeV.361202.Pythia8_Monash_NNPDF23LO_DD_minbias.recon.ESD.e3639_s2601_s2132_r6616
mc15_13TeV.361203.Pythia8_A2_MSTW2008LO_ND_minbias.evgen.EVNT.e3639
mc15_13TeV.361203.Pythia8_A2_MSTW2008LO_ND_minbias.recon.ESD.e3639_s2601_s2132_r6616
mc15_13TeV.361203.Pythia8_A2_MSTW2008LO_ND_minbias.recon.ESD.e3639_s2605_s2174_r6616
mc15_13TeV.361203.Pythia8_A2_MSTW2008LO_ND_minbias.recon.ESD.e3639_s2606_s2174_r6616
mc15_13TeV.361203.Pythia8_A2_MSTW2008LO_ND_minbias.recon.ESD.e3639_s2607_s2174_r6616
mc15_13TeV.361204.Pythia8_A2_MSTW2008LO_SD_minbias.evgen.EVNT.e3639
mc15_13TeV.361204.Pythia8_A2_MSTW2008LO_SD_minbias.recon.ESD.e3639_s2601_s2132_r6616
mc15_13TeV.361204.Pythia8_A2_MSTW2008LO_SD_minbias.recon.ESD.e3639_s2605_s2174_r6616
mc15_13TeV.361204.Pythia8_A2_MSTW2008LO_SD_minbias.recon.ESD.e3639_s2606_s2174_r6616
mc15_13TeV.361204.Pythia8_A2_MSTW2008LO_SD_minbias.recon.ESD.e3639_s2607_s2174_r6616
mc15_13TeV.361205.Pythia8_A2_MSTW2008LO_DD_minbias.evgen.EVNT.e3639
mc15_13TeV.361205.Pythia8_A2_MSTW2008LO_DD_minbias.recon.ESD.e3639_s2601_s2132_r6616
mc15_13TeV.361205.Pythia8_A2_MSTW2008LO_DD_minbias.recon.ESD.e3639_s2605_s2174_r6616
mc15_13TeV.361205.Pythia8_A2_MSTW2008LO_DD_minbias.recon.ESD.e3639_s2606_s2174_r6616
mc15_13TeV.361205.Pythia8_A2_MSTW2008LO_DD_minbias.recon.ESD.e3639_s2607_s2174_r6616
mc15_13TeV.361206.Pythia8_Monash_NNPDF23LO_DD_minbias_flux4_eps0085.evgen.EVNT.e3718
mc15_13TeV.361206.Pythia8_Monash_NNPDF23LO_DD_minbias_flux4_eps0085.recon.ESD.e3718_s2609_s2183_r6616
mc15_13TeV.361207.Pythia8_Monash_NNPDF23LO_DD_minbias_flux4_eps010.evgen.EVNT.e3718
mc15_13TeV.361207.Pythia8_Monash_NNPDF23LO_DD_minbias_flux4_eps010.evgen.EVNT.e3804
mc15_13TeV.361207.Pythia8_Monash_NNPDF23LO_DD_minbias_flux4_eps010.recon.ESD.e3804_s2609_s2183_r6616
mc15_13TeV.361208.Pythia8_Monash_NNPDF23LO_DD_minbias_flux5.evgen.EVNT.e3718
mc15_13TeV.361208.Pythia8_Monash_NNPDF23LO_DD_minbias_flux5.recon.ESD.e3718_s2609_s2183_r6616
mc15_13TeV.361209.Pythia8_Monash_NNPDF23LO_SD_minbias_flux4_eps006.evgen.EVNT.e3718
mc15_13TeV.361209.Pythia8_Monash_NNPDF23LO_SD_minbias_flux4_eps006.evgen.EVNT.e3804
mc15_13TeV.361209.Pythia8_Monash_NNPDF23LO_SD_minbias_flux4_eps006.recon.ESD.e3804_s2609_s2183_r6616
mc15_13TeV.361210.Pythia8_Monash_NNPDF23LO_SD_minbias_flux4_eps0085.evgen.EVNT.e3718
mc15_13TeV.361210.Pythia8_Monash_NNPDF23LO_SD_minbias_flux4_eps0085.recon.ESD.e3718_s2609_s2183_r6616
mc15_13TeV.361211.Pythia8_Monash_NNPDF23LO_SD_minbias_flux4_eps010.evgen.EVNT.e3804
mc15_13TeV.361211.Pythia8_Monash_NNPDF23LO_SD_minbias_flux4_eps010.recon.ESD.e3804_s2609_s2183_r6616
mc15_13TeV.361212.Pythia8_Monash_NNPDF23LO_SD_minbias_flux5.evgen.EVNT.e3718
mc15_13TeV.361212.Pythia8_Monash_NNPDF23LO_SD_minbias_flux5.recon.ESD.e3718_s2609_s2183_r6616
mc15_13TeV.361213.Pythia8_Monash_NNPDF23LO_DD_minbias_flux4_eps006.evgen.EVNT.e3726
mc15_13TeV.361213.Pythia8_Monash_NNPDF23LO_DD_minbias_flux4_eps006.evgen.EVNT.e3804
mc15_13TeV.361213.Pythia8_Monash_NNPDF23LO_DD_minbias_flux4_eps006.recon.ESD.e3804_s2609_s2183_r6616
mc15_13TeV.361223.Herwigpp_UEEE5_CTEQ6L1_MinBias.evgen.EVNT.e3907
mc15_13TeV.361224.Epos_minbias_inelastic.evgen.EVNT.e3908
mc15_13TeV.361224.Epos_minbias_inelastic.recon.ESD.e3908_s2601_s2174_r6616
mc15_13TeV.361225.Herwigpp_UEEE4_CTEQ6L1_MinBias.evgen.EVNT.e3930
mc15_13TeV.361234.Pythia8_Monash_NNPDF23LO_CD_minbias_flux5.evgen.EVNT.e3920
mc15_13TeV.361234.Pythia8_Monash_NNPDF23LO_CD_minbias_flux5.recon.ESD.e3920_s2609_s2183_r6616
mc15_13TeV.361235.QGSJet_minbias_inelastic.evgen.EVNT.e4405
mc15_13TeV.361235.QGSJet_minbias_inelastic.recon.ESD.e4076_s2601_s2174_r6616

Semantic Web with feedback