




Data Knowledge Base for HENP Experiments

Kurchatov Institute R&D Project

Maria Grigorieva

for NRC KI and TPU teams

Data Knowledge Base Kick-off meeting highlights. May 2016

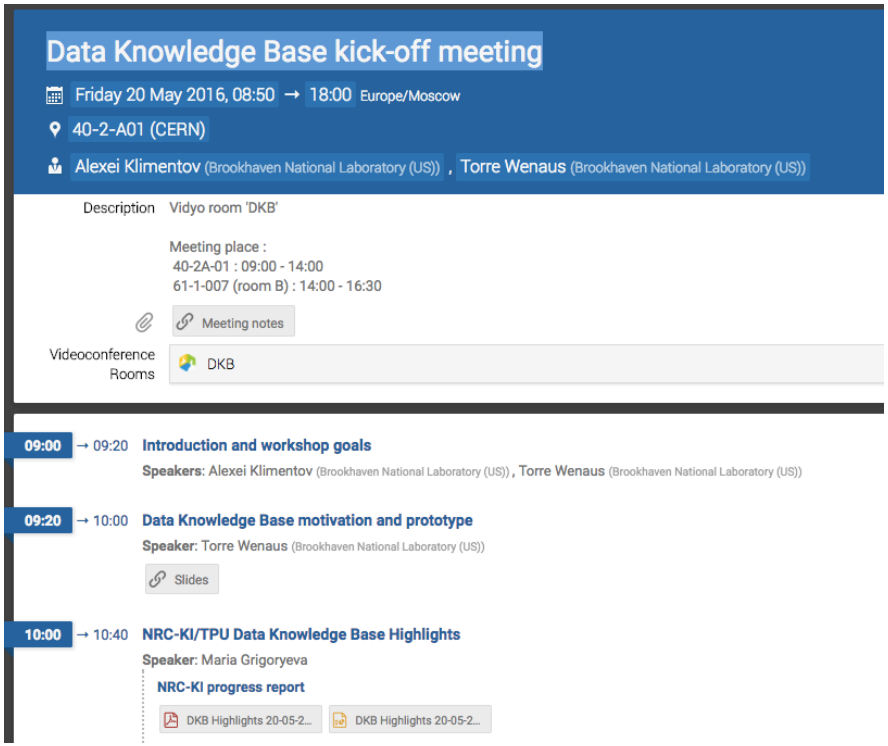
DKB Motivation

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

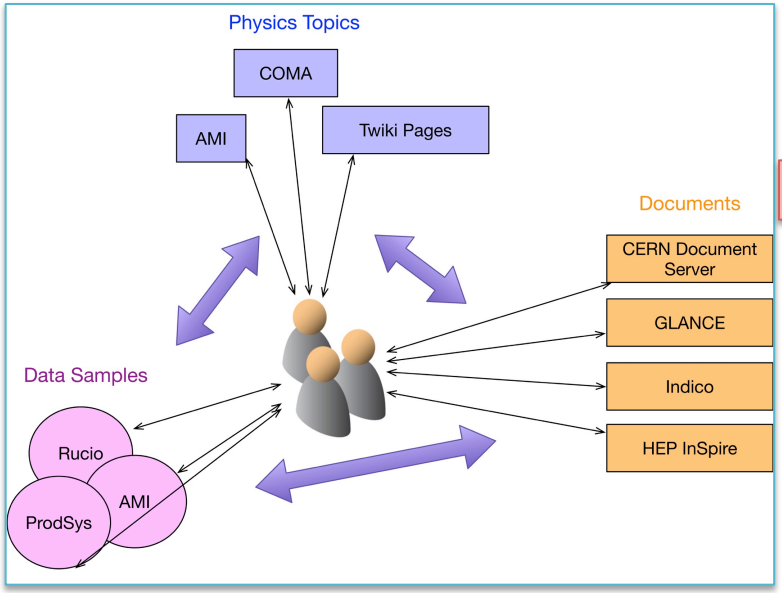
Organizing metadata in ATLAS, so as to provide a holistic view on physics topics, including integrated representation of all ATLAS documents (papers, drafts, supporting documents, conference notes, Indico meetings, Twiki pages, etc) and corresponding data samples (real data, MC datasets, containers).

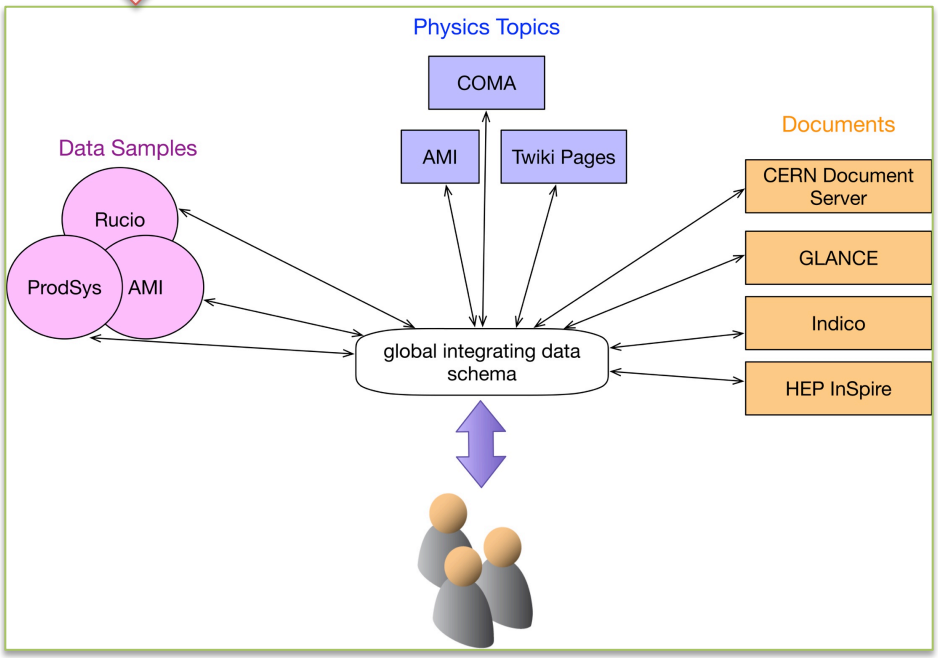


https://indico.cern.ch/event/527581/

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

DKB basic consideration





DKB R&D Topics

1. Ontological model of metadata
2. DKB Architecture
   - Setup and maintain of Virtuoso RDF Storage
   - Setup and maintain of Hadoop Transitional Storage
   - Consolidation of metadata in Virtuoso by means of Kafka Streams
   - Virtual data integration (with CDS, ProdSys)
   - Web Interface for Virtuoso RDF Storage
3. Data Processing
   - Modules for metadata export from GLANCE, CDS, ProdSys, AMI, for conversion to Turtle format and import in Virtuoso
   - PDF Analyzer (metadata extraction from PDF documents)

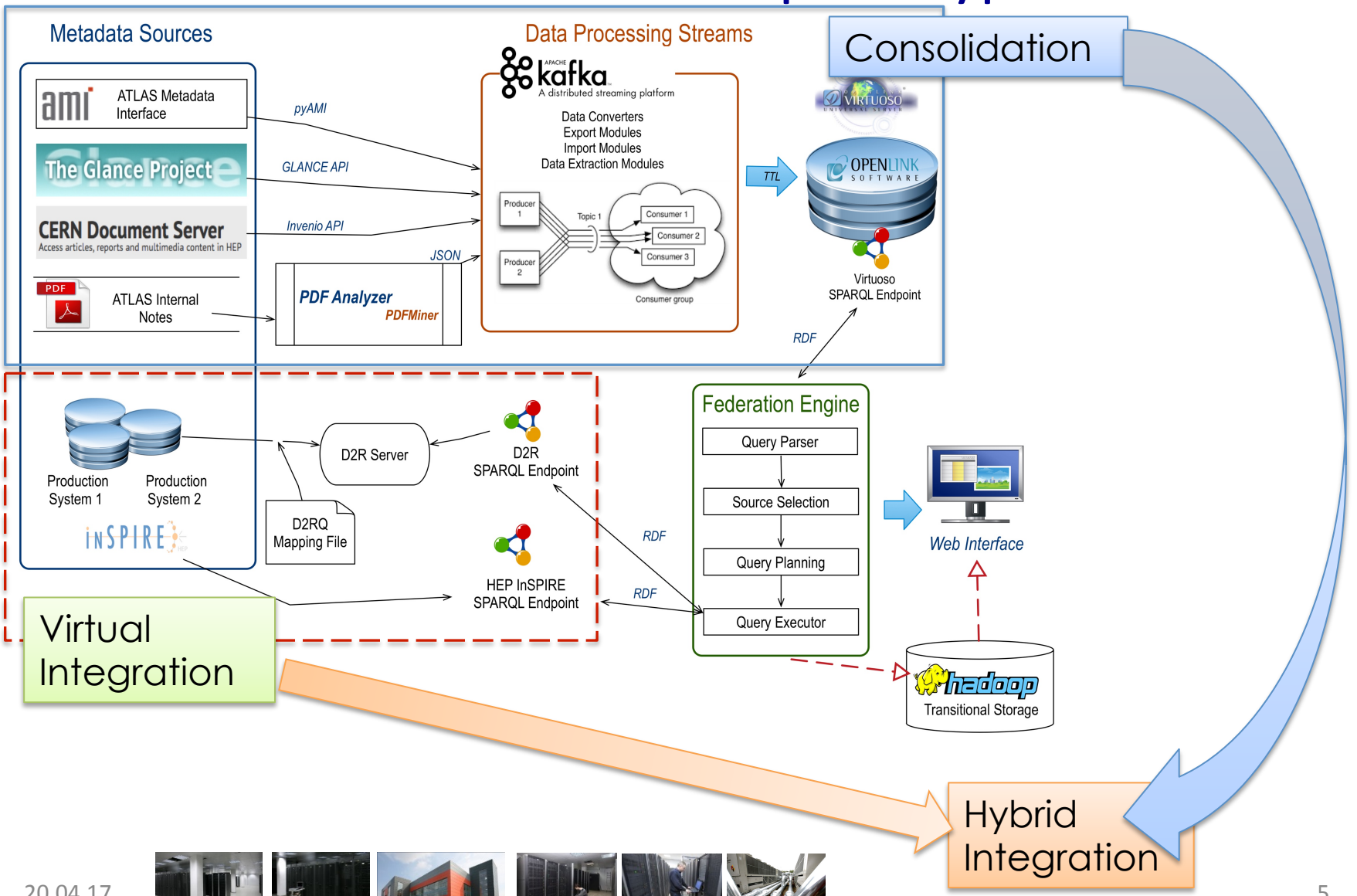
Tomsk Polytechnic University

Kurchatov Institute

1. **Maria Grigorieva (Team Lead)**
2. **Marina Golosova (core developer)**
3. Vasily Aulov
4. Maxim Gubin
5. Eugene Ryabinkin
6. Alexander Alexeev

NRC KI team qualification:
Development of Hybrid SQL/NoSQL PanDA Metadata Storage
- **https://indico.cern.ch/event/344958**

DKB architecture prototype



Data Integration Approaches

Metadata Consolidation captures data from multiple source systems and integrates it into a single persistent data store

+ Optimal performance and stability

- Requested data might be out of date because of the complexity of providiing data synchronization with various data sources

Metadata Federation provides a single virtual view of one or more data sources

+ Federated requests always return the actual data – data integrity support remains on the data source side

- The performance of federated queries depends on the communication channels and the queries execution rate on the side of data sources

Hybrid Metadata Integration:

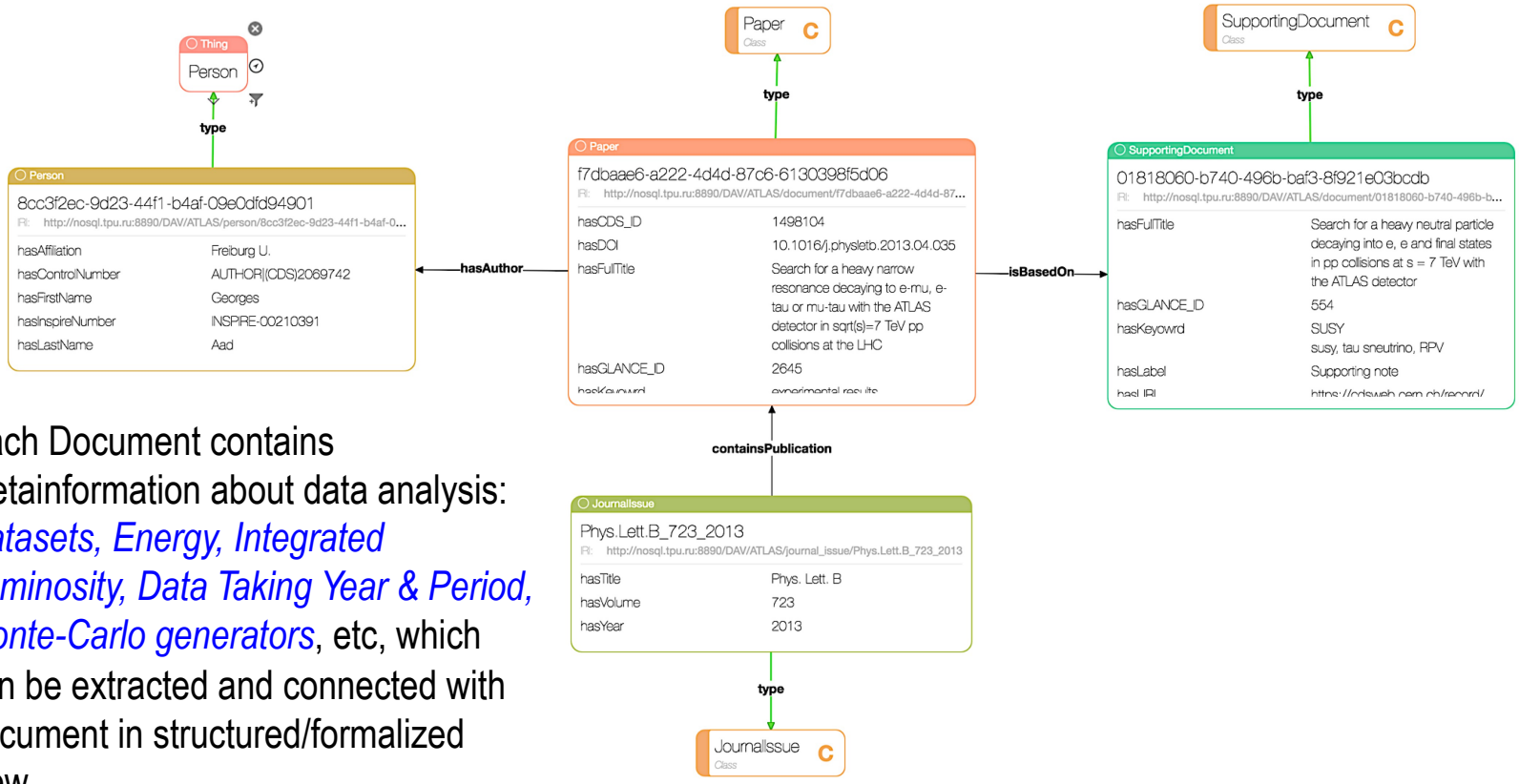
- Unchangeable data about global objects is consolidated:
  - Experiment Attributes
  - Links between documents
  - Documents general parameters from CDS
  - Metadata extracted from document's content
- Changing and auxiliary data is federated:
  - Detailed authors metadata from HEP InSPIRE system
  - Dataset's detailed metadata from ProdSys/AMI/Rucio

Hybrid Metadata Integration. Semantic Web

- *"The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation." // Berners-Lee, Scientific American, May 2001*
- Semantic Web consists primarily of three technical standards:
  - ***RDF (Resource Description Framework)***
  - ***SPARQL (SPARQL Protocol and RDF Query Language)***
  - ***OWL (Web Ontology Language)***
- RDF Statements are expressed in a "***triples***" <*subject, predicate, object*>.
- The entire universe can be described by triples because together, triples comprise a ***graph***.
- A graph can be linked to one or many other graphs on the World Wide Web and these graphs are a fundamental part of the Semantic Web.
- *Knowledge-oriented systems:* reasoning engines can be used to reason against assertions that have been made to infer new meaning, to find relationships and meaning far beyond the scope of the data, managed isolated.

ATLAS Metadata Ontological Model (1)

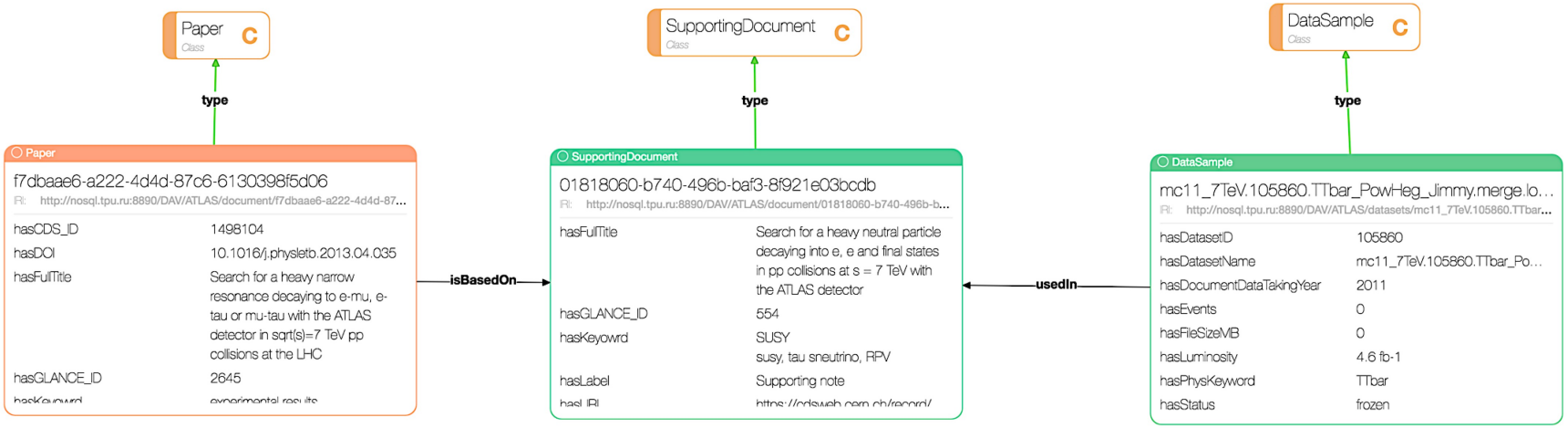
- ***Documents*** can be of different types.
- ***Scientific Paper*** is accompanied by ***Supporting Documents***.
- Document's inheritance is provided by "***isBasedOn***" Object Attribute.



- Each Document contains metainformation about data analysis: *Datasets, Energy, Integrated Luminosity, Data Taking Year & Period, Monte-Carlo generators*, etc, which can be extracted and connected with document in structured/formalized view.

ATLAS Metadata Ontological Model (2)

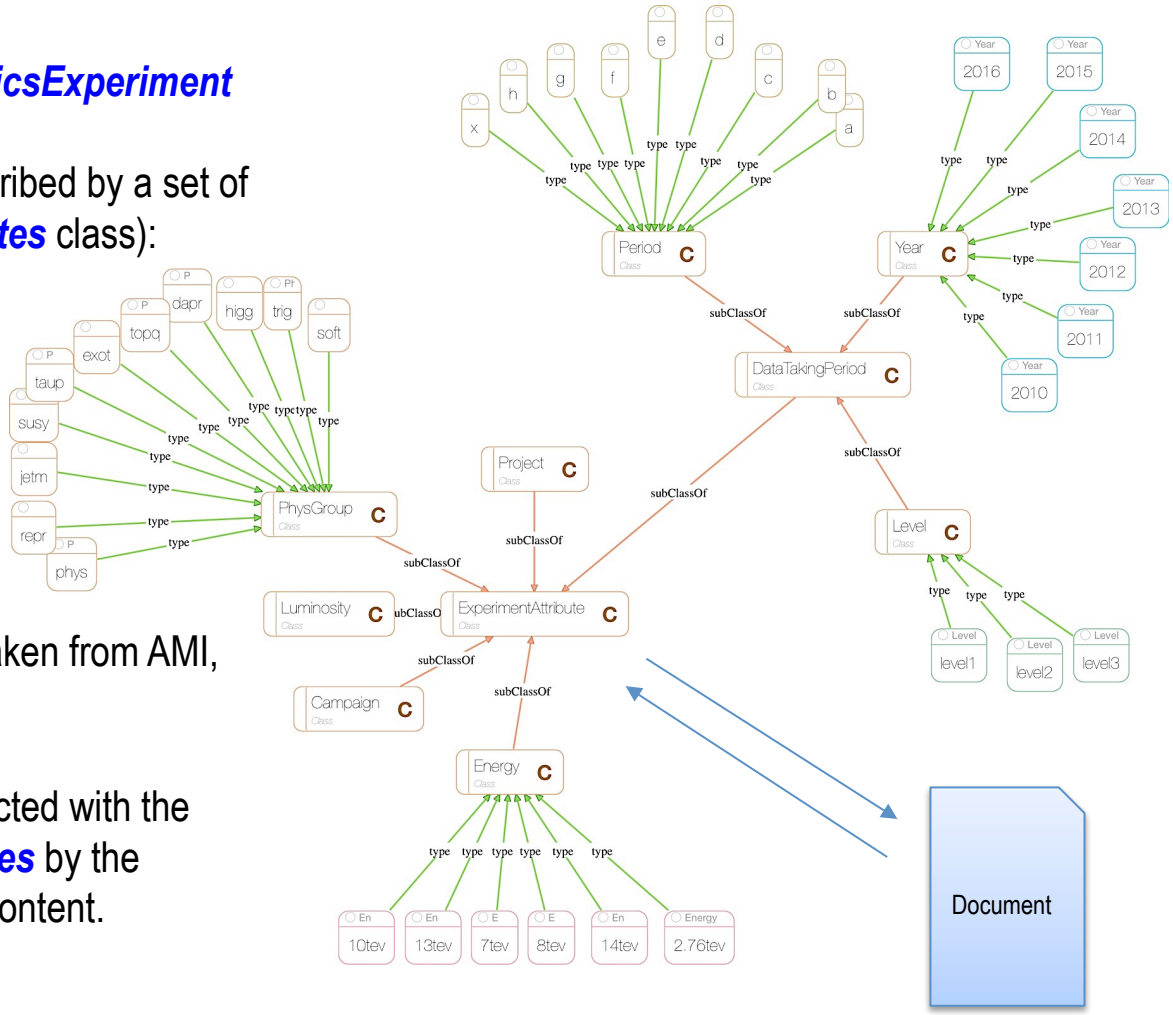
- ***DataSamples*** and ***Documents*** are connected by "***usedIn***"/"***referTo***" attributes.
- DataSample attributes are taken from ProductionSystem database:
  - hasDatasetID
  - hasStatus
  - hasEvents
  - hasTimestamp
  - ….
- In architecture prototypes v1 and v2 dataset's detailed metadata are consolidated in Virtuoso RDF-Storage.



ATLAS Metadata Ontological Model (3)

- Data Analysis in ATLAS = ***PhysicsExperiment*** class
- Each PhysicsExperiment is described by a set of parameters (***ExperimentAttributes*** class):

- Project *(ex: mc10_7TeV)*
- Campaign *(ex: mc11a)*
- Energy *(ex: 10TeV)*
- Integrated Luminosity
- Physics Group *(SUSY, HIGG,…)*
- Data Taking Period *(ex: 2010_A1_1)*
- *Other parameters are not defined yet*

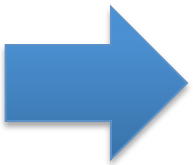


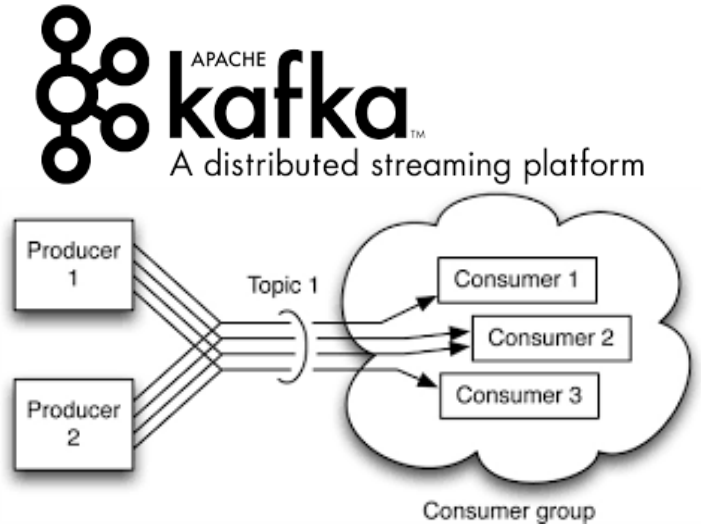
- ***ExperimentAttributes*** can be taken from AMI, Production System, Twiki pages.

- Each ***Document*** must be connected with the appropriate ***ExperimentAttributes*** by the parameters, extracted from the content.

Metadata Consolidation. Data Processing Agents

- GLANCE data processing
  - Export links between papers and supporting documents metadata from GLANCE in JSON
  - Convert links JSON to Turtle [RDF syntax - https://www.w3.org/TeamSubmission/turtle/]
  - Import links to Virtuoso module
- CDS data processing
  - Export paper's metadata from CDS
  - Convert Paper's JSON to Turtle
  - Import Paper's Turtle to Virtuoso
  - Export supporting notes metadata from CDS
  - Convert supporting notes JSON to Turtle
  - Import supporting notes Turtle to Virtuoso
  - Get URL of PDF documents (supporting documents)
  - Downloading PDF Documents
  - PDF Analyzer module
  - Converting PDF Analyzer results from JSON to Turtle
  - Import results to Virtuoso

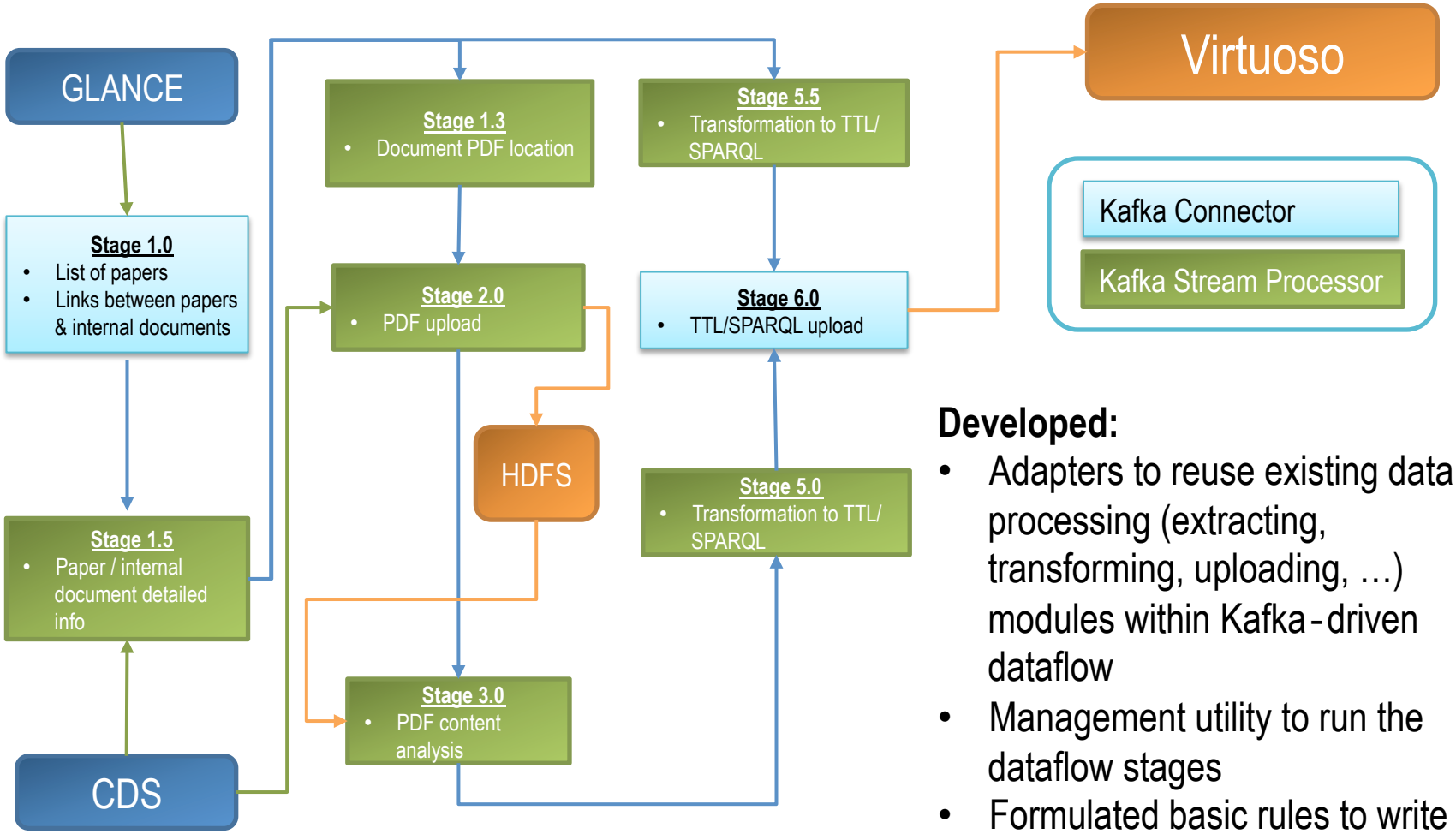




Dataflow development issues

- Data transfer from "upstream" to "downstream" stages
- Data preservation between stages
- Obsolete data removal
- Guarantee of reprocessing on data/ process update
- Processing delay
- Failure recovery

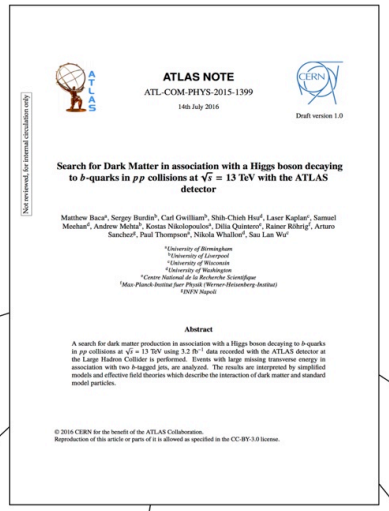
Kafka Streams for dataflow automation



Developed:

- 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 the text of unstructured documents



ATLAS NOTE
ATL-COM-PHYS-2015-1399
14th July 2016
Draft version 1.0

Search for Dark Matter in association with a Higgs boson decaying to b-quarks in pp collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector

Abstract

A search for dark matter production in association with a Higgs boson decaying to b-quarks in pp collisions at $\sqrt{s}$ = 13 TeV using 3.2 fb^{-1} data recorded with the ATLAS detector at the Large Hadron Collider is performed. Events with large missing transverse energy in association with two b-tagged jets, are analyzed. The results are interpreted by simplified models and effective field theories which describe the interaction of dark matter and standard model particles.

hasContent

GRL used in this analysis is DATA15_13TEV.PERIODALLYEAR_DETSTATUS-V73-PRO19-08_DQDEFECTS-00-01-02_PHYS_STANDARDGRL_ALL_GOOD_25NS.XML

hasContent

Abstract

A search for dark matter production in association with a Higgs boson decaying to b-quarks in pp collisions at $\sqrt{s}$ = 13 TeV using 3.2 fb^{-1} data recorded with the ATLAS detector at the Large Hadron Collider is performed. Events with large missing transverse energy in association with two b-tagged jets, are analyzed. The results are interpreted by simplified models and effective field theories which describe the interaction of dark matter and standard model particles.

hasContent

DSID	Sample Name	Tag
341100	Pythia8EvtGen_A14NNPDF23LO_WlvH125_bb	e3885_s2608_s2183_r6869_r6282_p2419
341101	Pythia8EvtGen_A14NNPDF23LO_ZvvH125_bb	e3885_s2608_s2183_r6869_r6282_p2419
341102	Pythia8EvtGen_A14NNPDF23LO_ZllH125_bb	e3885_s2608_s2183_r6869_r6282_p2419

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

hasContent

hasContent

L Monte Carlo Samples

PYTHIA **6.423**

mc10_7TeV.105009.J0_pythia_jetjet.merge.NTUP_BTAG.e574_s934_s946_r1653_r1700_p370
..
mc10_7TeV.105016.J7_pythia_jetjet.merge.NTUP_BTAG.e574_s934_s946_r1653_r1700_p370

PYTHIA **6.423+**PILEUP

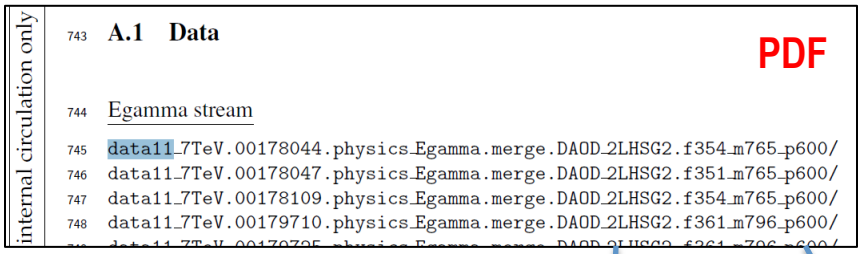
mc10_7TeV.105009.J0_pythia_jetjet.merge.NTUP_BTAG.e574_s934_s946_r1833_r1700_p370
..
mc10_7TeV.105016.J7_pythia_jetjet.merge.NTUP_BTAG.e574_s934_s946_r1833_r1700_p370

hasContent

Real datasets

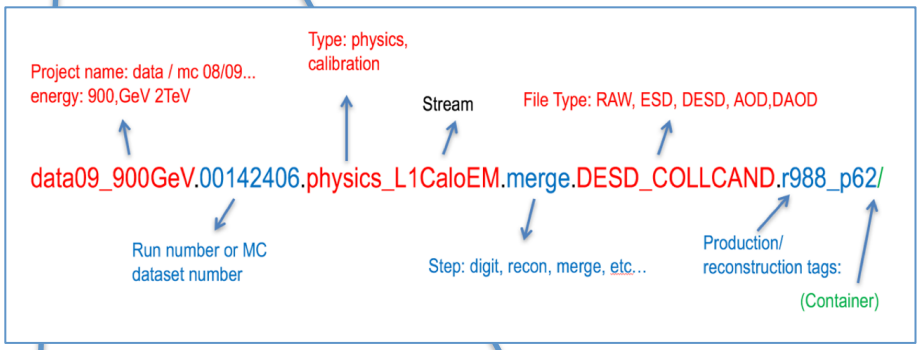
DSID	Sample Name	Tag
00279598	physics_Main	f628_m1497_p2425
00279685	physics_Main	f628_m1497_p2425
00279764	physics_Main	f628_m1497_p2425
00279813	physics_Main	f628_m1497_p2425
00279867	physics_Main	f628_m1497_p2425
00279928	physics_Main	f628_m1497_p2425
00279932	physics_Main	f629_m1504_p2425
00279984	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

Datasets extraction from PDF documents



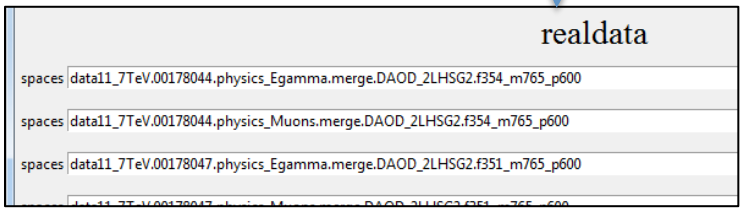
Regular expressions are constructed according to ATLAS dataset nomenclature. These expressions are used to extract the dataset names from the text.

PDFMiner is a tool for extracting information from PDF documents.



GUI interface:

- Edit dataset names
- Delete datasets from the list
- Export resulted list to JSON file



JSON

{ "content": { "real_datasets": ["data11_7TeV.
00178044.physics_Egamma.merge.DAOD_2LHSG2.f354_m765_p600",
"data11_7TeV.00178047.physics_Egamma.merge.DAOD_2LHSG2.f351_m765_p600",
"data11_7TeV.00178109.physics_Egamma.merge.DAOD_2LHSG2.f354_m765_p600",
"data11_7TeV.00179710.physics_Egamma.merge.DAOD_2LHSG2.f361_m796_p600",
"data11_7TeV.00179725.physics_Egamma.merge.DAOD_2LHSG2.f361_m796_p600",
.....

Extraction of data from tables

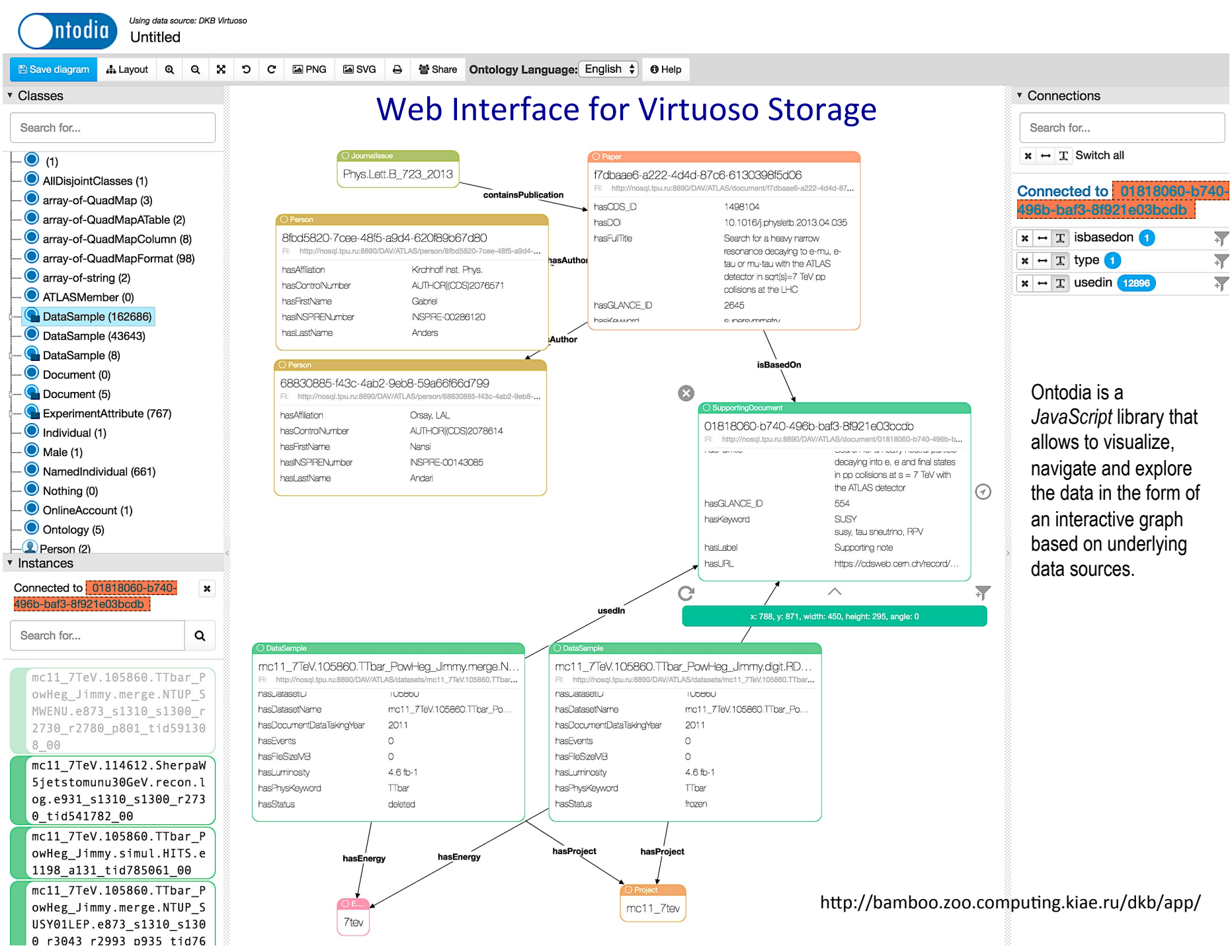
Signal Point $M(\tilde{g})$ [GeV]	Signal Point $M(\tilde{\chi}_1^0)$ [GeV]	Run Number	Cross Section (LO) [pb]	Signal Point $M(\tilde{g})$ [GeV]	Signal Point $M(\tilde{\chi}_1^0)$ [GeV]	Run Number	Cross Section (LO) [pb]
400	50	123078	6.00	900	50	138568	6.19×10^{-3}
400	75	123079	5.95	900	100	138569	6.14×10^{-3}
400	100	123080	6.00	900	150	138570	6.10×10^{-3}
400	125	123081	6.02	900	200	138571	6.08×10^{-3}
400	150	118430	6.03	900	300	138572	5.92×10^{-3}

Table 26: GGM signal samples. Each signal point is defined by the gluino and lightest neutralino mass. The run number and LO cross section is given.

```
<?xml version="1.0" encoding="utf-8" ?>
<pages>
  <page id="1" bbox="0.000,0.000,612.000,792.000" rotate="0">
    <textbox id="0" bbox="229.080,675.825,366.146,691.074">
    <textline bbox="229.080,675.825,366.146,691.074">
      <text font="MGTNRE+CMSSBX10" bbox="229.080,675.825,244.226,691.074" size="15.249">A</text>
      <text font="MGTNRE+CMSSBX10" bbox="242.273,675.825,257.419,691.074" size="15.249">T</text>
      <text font="MGTNRE+CMSSBX10" bbox="257.386,675.825,269.391,691.074" size="15.249">L</text>
      <text font="MGTNRE+CMSSBX10" bbox="269.396,675.825,284.541,691.074" size="15.249">A</text>
      <text font="MGTNRE+CMSSBX10" bbox="284.509,675.825,297.134,691.074" size="15.249">S</text>
    </textline>
    </textbox>
    <figure name="R5" bbox="70.920,642.800,155.880,750.800">
      <image width="84" height="108" />
    </figure>
    <layout>
      <textgroup bbox="73.080,185.033,522.095,691.074">
      <textgroup bbox="73.080,380.838,522.095,691.074">
    </layout>
  </page>
```

Tables are found by looking for their descriptions in text. After finding the description, the page containing the table is extracted into XML format and the table is reconstructed using the information from it.

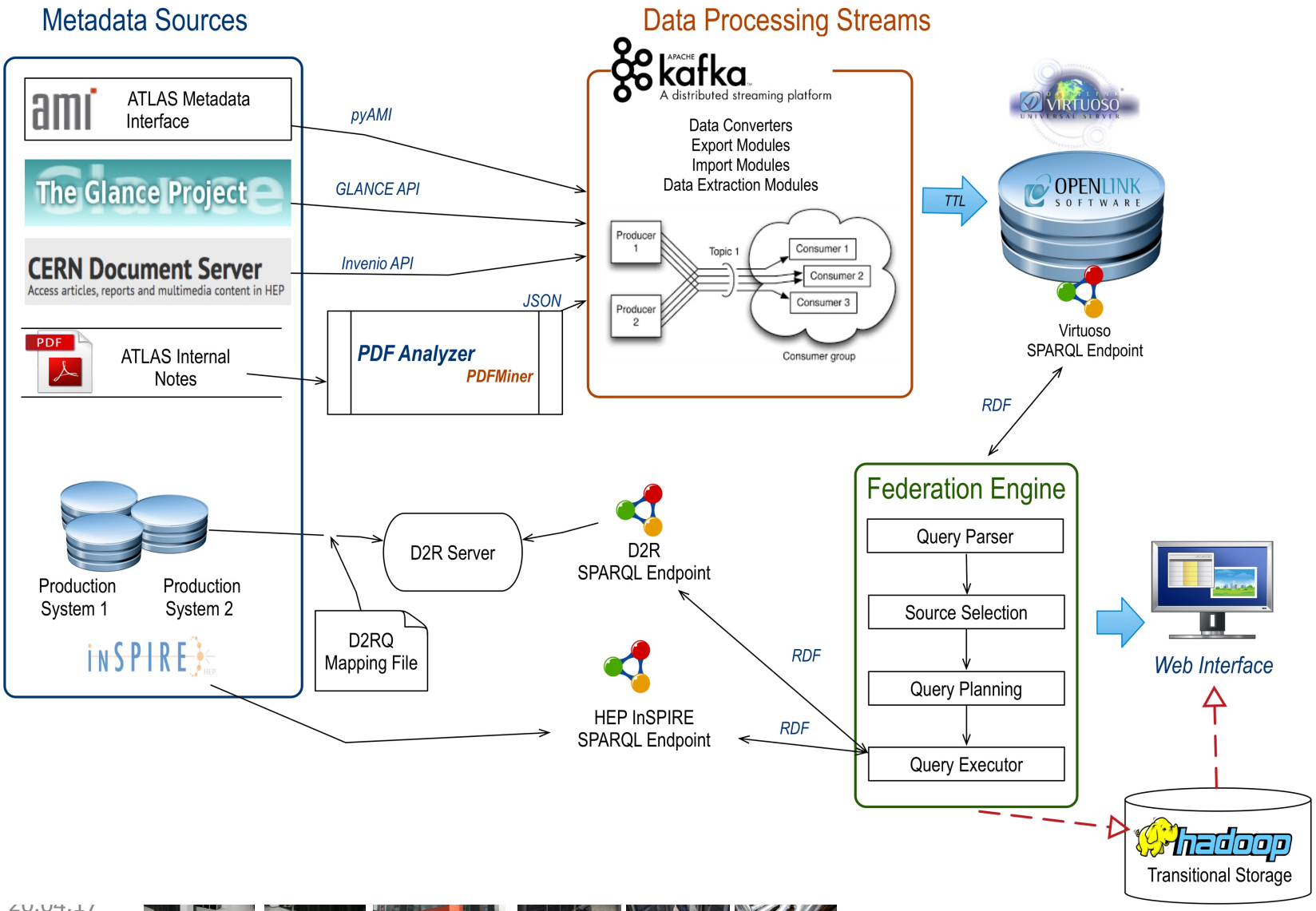

```
"table_26": [
   "Table 26: GGM signal samples. Each signal point i
   [
      [
         "Signal",
         "Point",
         "Run Number",
         "Cross Section (LO)",
         "Signal",
         "Point",
         "Run Number",
         "Cross Section (LO)"
      ],
      [
         "M( \u02dcg)",
         "M( \u02dc\u03c701)",
         "EMPTY",
         "[pb]",
         "M( \u02dcg)",
         "M( \u02dc\u03c701)",
         "EMPTY",
         "[pb]"
      ],
      [
         "[GeV]",
         "[GeV]",
         "EMPTY",
         "EMPTY",
         "[GeV]",
         "[GeV]",
         "EMPTY",
         "EMPTY"
      ],
      [
         "400",
         "50",
         "123078",
         "6.00",
         "900",
         "50",
         "138568",
         "6.19\u00d710\u22123"
      ],
      [
         "400",
         "75",
         "123079",
         "5.95",
         "900",
         "100",
         "138569",
         "6.14\u00d710\u22123"
      ],
```

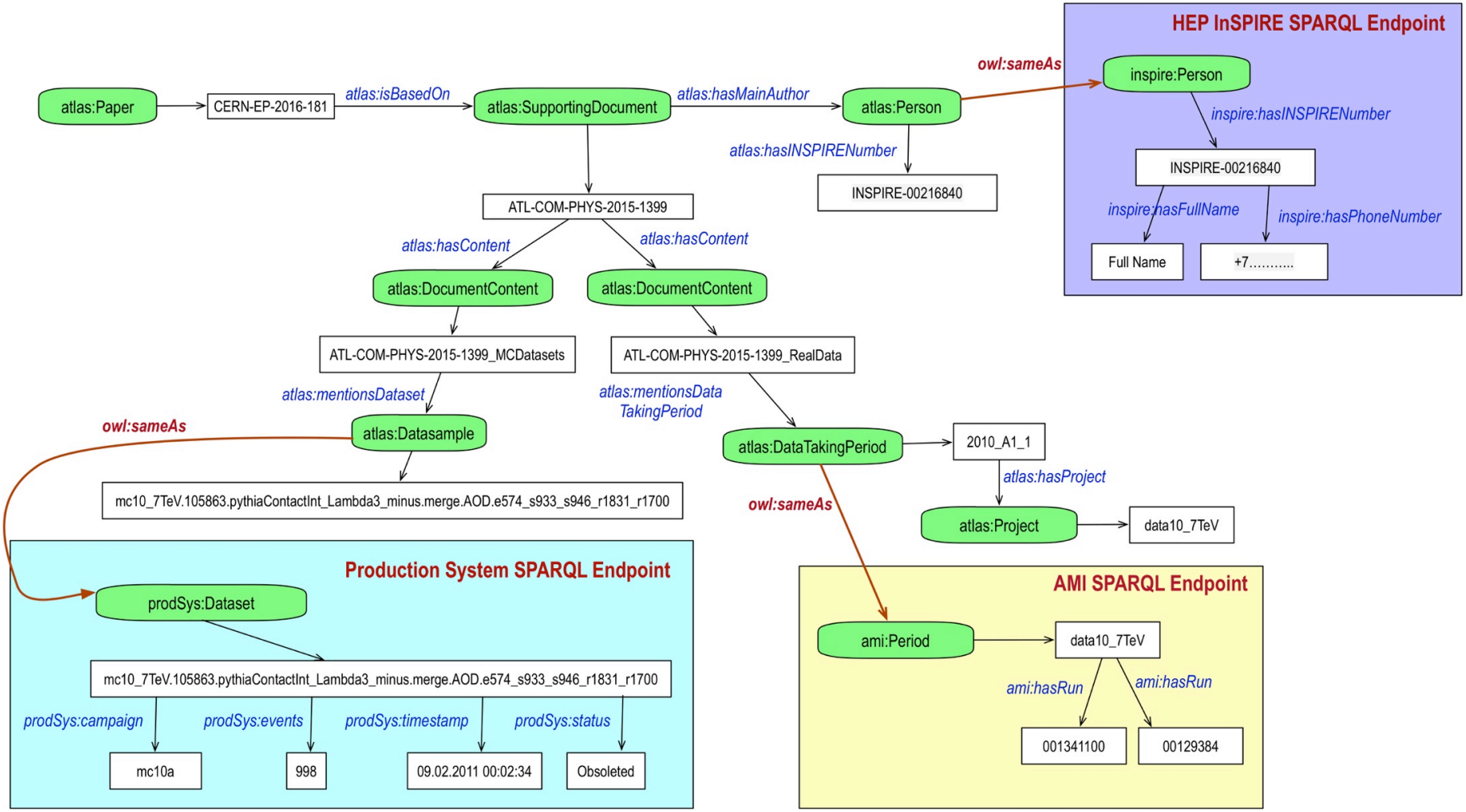
Ontodia is a *JavaScript* library that allows to visualize, navigate and explore the data in the form of an interactive graph based on underlying data sources.

http://bamboo.zoo.computing.kiae.ru/dkb/app/

DKB architecture prototype



Federated Requests Example



Summary, Outline and Future Plans

- The team from Kurchatov Institute and TPU worked on DKB prototype and tools evaluation
- During the first year the Data Knowledge Base prototype was designed and implemented
  - we use hybrid integration approach (*consolidation* + *virtual integration*)
  - automated data flows processing (using Apache Kafka)
  - we provide a user friendly Web I/F. It allows to present metadata cross-references in the form of graphs
  - we designed a global ontological data schema, to represent Documents, Physics Topic and Data Samples metadata in a coherent way
  - we evaluated tools and technologies and made a choice of
    - virtuoso, apache kafka, ontodia, …
  - code repository is in SVN
- Program to extract metadata information from unstructured scientific documents in PDF format was developed, coded and implemented

Summary, Outline and Future Plans. Cont'd

- The following matters will be addressed in a near-term future
  - code repository migration to github
  - DKB access authentication
    - the most probable candidate is CERN SSO
      - implemented for BigPanDA monitor and ProdSys
  - Virtuoso scalability studies
  - future development of a global metadata integration model, based on hybrid data integration approach
  - Refining Kafka Streams dataflows automation
  - Enhancing semi-automatic PDF Analyzer functionality:
    - improve GUI and add user's option to edit automatic meta-data extraction results

Possible Contribution to the ATLAS DCC project.

Dataset discovery and 'whiteboard' sub-project(s)

- Development of the ontological data models for various sources of ATLAS metadata
- Execution of the consolidation dataflow, based on metadata from GLANCE, CDS, AMI using PDF Analyzer
  - as a component in addition to existing (or/and planned to be developed tools)
- The choice of technology and implementation of a SPARQL endpoint for metainformation from
  - ProdSys1&2 / Rucio / AMI (?), HEP InSPIRE / CDS (?)
  - Execution of the federated SPARQL requests (just examples):
    - get all datasets and related meta information from ProdSys/AMI/Rucio if they found in ATLAS papers or/and in Supporting Documents
    - retrieve a list of documents referred to data analysis conducted using data from period *X* and year *Y*
    - retrieve all Documents for a specific physics group, published in year *XYZ* and reference to produced datasets and datasets states
    - get detailed information about **main** authors for Paper from InSPIRE and return titles of related ATLAS publications
    - retrieve all documents where specific data sample is mentioned, with detailed metadata about this dataset from ProdSys, and find metadata about main authors of this document in InSPIRE

Findings and Questions :

- It is not always clear what is the best source to find information about
  - ATLAS projects
  - Campaigns and sub-campaigns
    - and description
  - Data taking periods

It would be beneficial to have above descriptions in more formalized format

Pointing to the source of information to be used for our studies will help

- Datasets (and data samples) are described in several places
  - We didn't find one with the complete meta-data info
    - AMI, Rucio, ProdSys : each has a part of info
      - » we didn't study coherency between them, just an observation
- Authors & Publications
  - CDS/InSPIRE (?)

Findings and Questions (cont'd) :

- What is a combination of meta-information or/and parameters to identify the data sample used for a particular physics analysis :
  - Project(s)
  - Campaign(s) and sub-campaign(s)
  - integrated luminosity (and statistics)
  - Physics Groups
  - Data taking period(s)
  - SW release

  ....

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

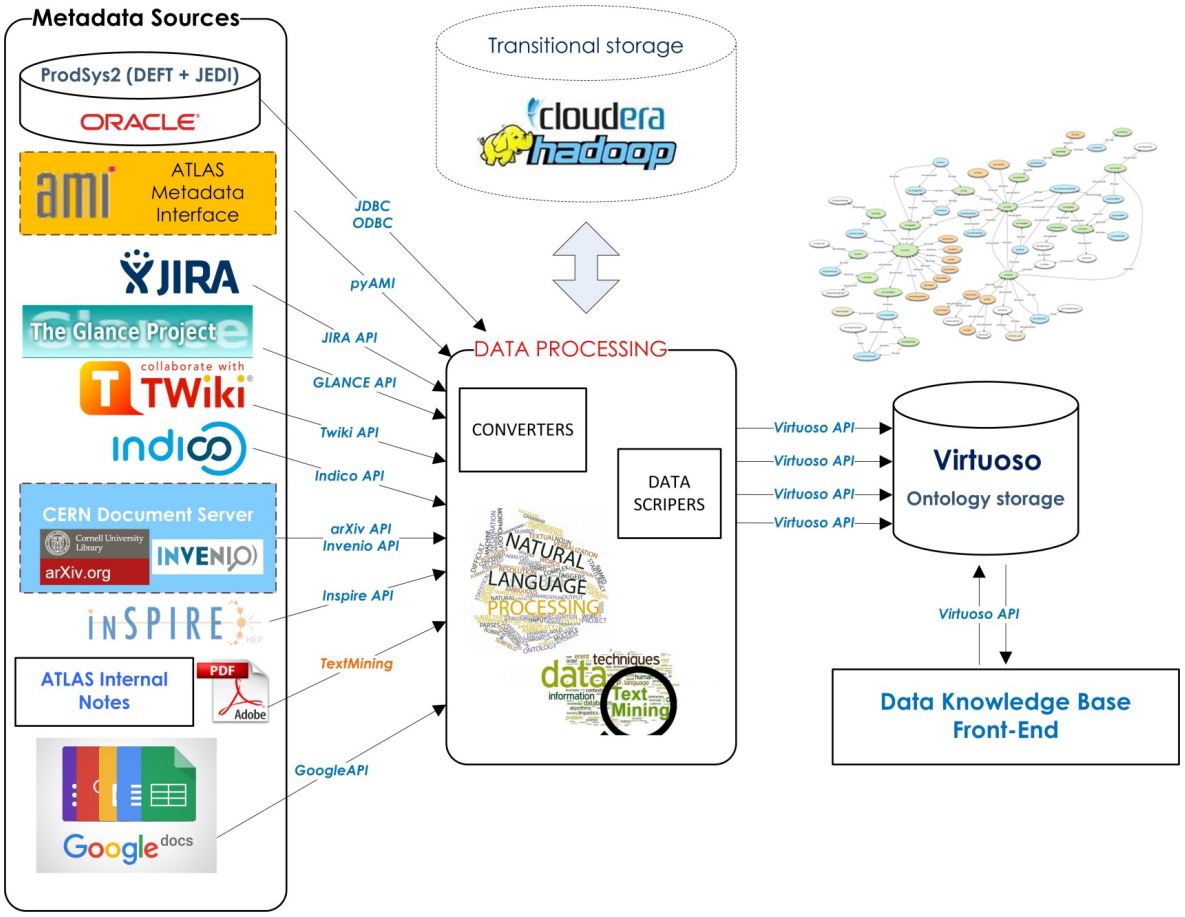
the Russian Foundation for Basic research under contract #16-37-00246.






ADDITIONAL SLIDES

DKB architecture prototype v.1



Ontology storage – OpenLink Virtuoso:

- Developed first prototype of the ontology for ATLAS Data Analysis.
- Virtuoso ontology storage installed in Tomsk Polytechnic University

Transitional Hadoop Storage in Kurchatov Institute

- Production System metadata [datasets] was exported from Oracle DB and imported to Hadoop Storage

ATLAS Internal Notes processing:

- Developed PDF Analyzer tool, based on PDFMiner, which extracts dataset names from full texts of ATLAS Internal Notes

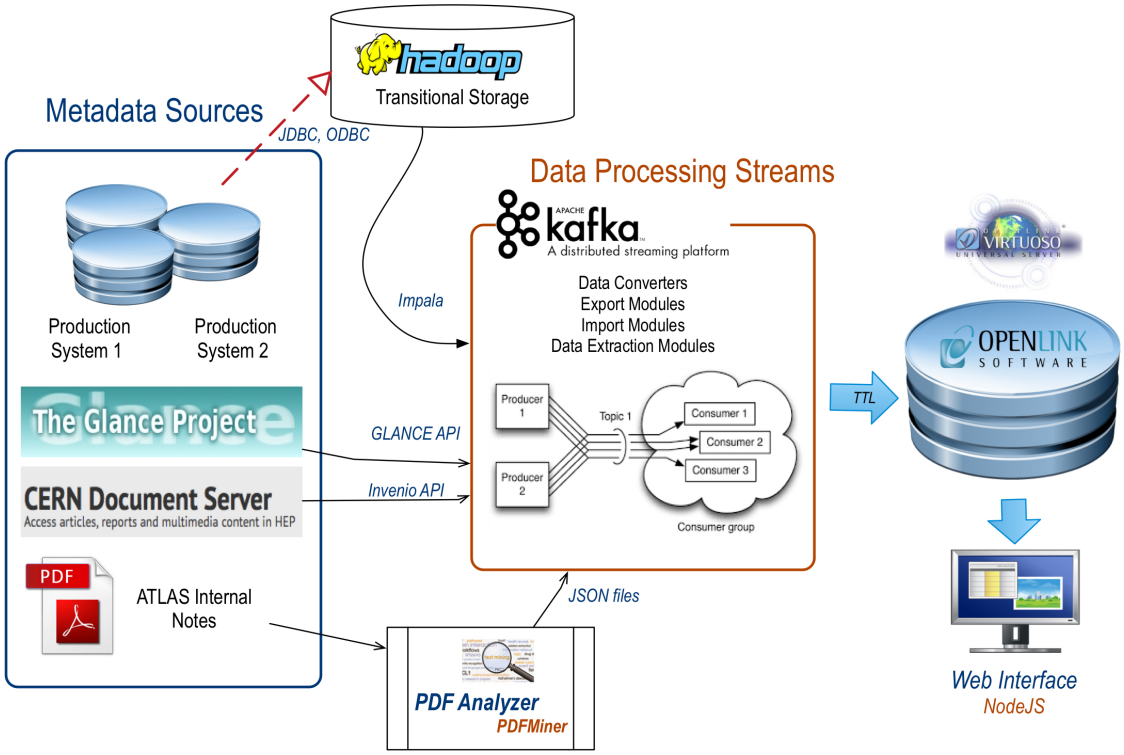
Data Processing:

- Developed tools, converting the metadata from GLANCE, CERN Document Server and Production System [datasets] to TTL format for Virtuoso storage
- Metadata consolidated in Virtuoso storage

DKB architecture prototype v.2

Data Processing:

- Implemented Metadata Integration Chain using data streaming based on Apache Kafka to automate data processing workflows
- Executed test data flow which export metadata from GLANCE and CDS, and import integrated metadata in TTL format in Virtuoso Storage



Enhanced PDF Analyzer functionality:

- Extract metadata from PDF Tables

Enhanced ontological model with metadata from AMI:

- Data taking periods
- Run Numbers
- Projects
- Campaigns

Virtuoso Server = http://nosql.tpu.ru:8890/

SCHEMA GRAPH:
http://nosql.tpu.ru:8890/DAV/home/dba/ATLAS

RESOURSE GRAPH:
http://nosql.tpu.ru:8890/DAV/ATLAS

Subject	Predicate	Object
Document/CERN-EP-2016-181	hasType	Paper
Document/CERN-EP-2016-182	hasTitle	Search for dark matter in association with a Higgs boson decaying to b-quarks in pp collisions at sv=13 TeV with the ATLAS detector
Document/CERN-EP-2016-183	hasArXiv	arXiv:1609.04572
Document/CERN-EP-2016-184	hasKeyword	exotics
Document/CERN-EP-2016-185	isBasedOn	Document/ATL-COM-PHYS-2015-1399
Document/ATL-COM-PHYS-2015-1399	hasType	SupportingDocument
Document/ATL-COM-PHYS-2015-1400	hasTitle	Search for dark matter in association with a Higgs boson decaying to b-quarks in pp collisions at sv=13 TeV with the ATLAS detector
Document/ATL-COM-PHYS-2015-1401	useGRL	data15_13TeV.periodAllYear_DetStatus-v73-pro19-08_DQDefects- 00-01-02_PHYS_StandardGRL_All_Good_25ns.xml
Document/ATL-COM-PHYS-2015-1402	hasEnergy	13TeV
Document/ATL-COM-PHYS-2015-1403	hasPublicationYear	2016
Document/ATL-COM-PHYS-2015-1404	hasContent	
Document/ATL-COM-PHYS-2015-1405		Content/ATL-COM-PHYS-2015-1399_Table_19
Content/ATL-COM-PHYS-2015-1399_Table_19	hasDescription	Monte Carlo samples used as baseline for Standard Model VH(→ bb)
Content/ATL-COM-PHYS-2015-1399_Table_20	mentionsDataSample	DataSample/ATL-COM-PHYS-2015-1399_341100
Content/ATL-COM-PHYS-2015-1399_Table_21	mentionsDataSample	DataSample/ATL-COM-PHYS-2015-1399_341101
Content/ATL-COM-PHYS-2015-1399_Table_22	mentionsDataSample	DataSample/ATL-COM-PHYS-2015-1399_341102
DataSample/ATL-COM-PHYS-2015-1399_341100	hasType	MC
DataSample/ATL-COM-PHYS-2015-1399_341100	hasProject	MC15_13TeV
DataSample/ATL-COM-PHYS-2015-1399_341100	hasDataSampleID	341100
DataSample/ATL-COM-PHYS-2015-1399_341100	hasSampleName	Pythia8EvtGen_A14NNPDF23LO_WlvH125_bb
DataSample/ATL-COM-PHYS-2015-1399_341100	hasTag	e3885_s2608_s2183_r6869_r6282_p2419

SPARQL Endpoint:
http://nosql.tpu.ru:8890/sparql

Reconstruction of the real and MC datasets from document content

B Signal MC Samples

Table 21 shows a full list of signal MC samples of MC11a used in the council conf note. For the paper, MC11b is used. The difference on AMI tag is `r2920_r2900_p756` instead of `r2730_r2700_p756`.

Table 21: List of signal MC samples used (MC11b production).

	Process	Dataset ID	Generator	Filter	AMI tag
m_H=100 GeV	$gg \to H$	116866	PowHeg	-	e873_s1310_s1300_r2920_r2900_p756
	VBF	125170	PowHeg	-	e893_s1310_s1300_r2920_r2900_p756
	WH	125329	Pythia	-	e825_s1310_s1300_r2920_r2900_p756
	ZH	125489	Pythia	-	e825_s1310_s1300_r2920_r2900_p756
	$t\bar{t}H$	116064	Pythia	PhotonFilter	e893_s1310_s1300_r2920_r2900_p756
m_H=105 GeV	$gg \to H$	116867	PowHeg	-	e873_s1310_s1300_r2920_r2900_p756
	VBF	125171	PowHeg	-	e893_s1310_s1300_r2920_r2900_p756
	WH	125330	Pythia	-	e825_s1310_s1300_r2920_r2900_p756
	ZH	125490	Pythia	-	e825_s1310_s1300_r2920_r2900_p756
	$t\bar{t}H$	116065	Pythia	PhotonFilter	e825_s1310_s1300_r2920_r2900_p756

Not unique values:
could repeat in
different projects

Campaign Description from AMI

```
…
}, {
   "MC11a": {
      "mc11_7TeV": {
         "digit": {
            "RDO": ["d579", "d580"]
         },
         "merge": {
            "AOD": ["r2700", "r2780"]
         },
         "recon": {
            "*": ["a128", "a131", "a133",
"a134", ...]
         }
      }
   }
},
…
```

Project

mc11_7TeV.116866.*.*.*.e873_s1310_s1300_r2920_r2900_p756
mc11_7TeV.125170.*.*.*.e893_s1310_s1300_r2920_r2900_p756
mc11_7TeV.125329.*.*.*.e825_s1310_s1300_r2920_r2900_p756
mc11_7TeV.125489.*.*.*.e825_s1310_s1300_r2920_r2900_p756
mc11_7TeV.116064.*.*.*.e893_s1310_s1300_r2920_r2900_p756

Reconstruction of the real and MC datasets from document content

Measurement of underlying event characteristics using charged particle jets in pp collisions at $\sqrt{s} = 7$TeV with the ATLAS detector at the LHC

All data used in this analysis were taken during the 2010 LHC running period A (run numbers 152166-153200) and period B (run numbers 153565-155160), with the May reprocessing (release 16).

Project Name
"data10_7TeV"

Period's Description from AMI


```
[{"period": "A",
  "periodLevel": "2",
  "projectName": "data10_7TeV",
  "description": "unsqueezed stable beam data
(beta*=10m): typical beam spot width in x and
y is 50-60 microns.",
  "status": "locked"},
{"period": "B",
 "periodLevel": "2",
 "projectName": "data10_7TeV",
 "description": "first squeezed stable beams
(beta*=2m): typical beam spot width in x and y
is 30-40 microns.",
 "status": "locked"}, {
```


Real data

generator	configuration	sample number
Pythia 6	AUET2B	126169, 126346-126349
Pythia 6	Z1	126172, 126358-126361
Pythia 6	Perugia2011	126170, 126354-126357
Pythia 6	Perugia2011 NOCR	126171, 126350-126353
Herwig++	UE7000	113906-113909
Pythia 8.145	4C	108316-108318, 108351, 113118-113125

Table 2: EVGEN Monte Carlo samples

underlying physics distributions by using different MC *tunes* (configurations) as control samples. We use Pythia 6 (MC09) as the primary control sample because it has the best available statistics (19.6M events) compared to the data (42.6M events). More information about the MC samples can be found in Table 1.

Project Name
"mc09_7TeV"

MC

ATLAS NOTE

ATL-COM-PHYS-2010-685

March 9, 2011

$W \to \mu\nu$ and $Z \to \mu\mu$ cross-sections measurements in proton-proton collisions at $\sqrt{s} = 7$ TeV with the ATLAS Detector

Abstract

This paper presents the measurements of W, W^+ and W^- to muon and $Z \to \mu^+\mu^-$ inclusive cross-sections with the ATLAS detector in proton-proton collisions at $\sqrt{s}$=7 TeV. The results presented are based on an integrated luminosity of 310 nb^{-1} for the W analysis and 331 nb^{-1} for the Z analysis, collected in April-July 2010 with fully operational detector and stable beam conditions. There are 1181 W and 109 Z candidate events in the muon decay channel. The distributions for the main observables are compared to a PYTHIA Monte Carlo simulation at different stages of the selection. We measure $\sigma_W \times BR(W \to \mu\nu) = 9.58 \pm 0.30(\text{stat}) \pm 0.50(\text{sys}) \pm 1.05(\text{lum})$ nb and $\sigma_{Z/\gamma^*} \times BR(Z/\gamma^* \to \mu^+\mu^-) = 0.87 \pm 0.08(\text{stat}) \pm 0.05(\text{sys}) \pm 0.10(\text{lum})$ nb, consistent with the Standard Model expectations.

Process	Dataset	Generator	Cross-section (pb)	N_{evt} ($\times 10^6$)	note
$W \to \mu\nu$	106044	PYTHIA	10 454	7	
$Z \to \mu\mu$	106047	PYTHIA	989	7.9	$\sqrt{\hat{s}} > 60$ GeV
$W \to \tau\nu$	106022	PYTHIA	10 454	5	single lepton filter ($\varepsilon = 0.877$) times lepton branching ratio
$Z \to \tau\tau$	106052	PYTHIA	989	7.9	$\sqrt{\hat{s}} > 60$ GeV
$t\bar{t}$	105861	POWHEG	161	0.2	$m_t = 172.5$ GeV/c^2, single lepton filter $\varepsilon = 0.538$
$b\bar{b}$	108405	PYTHIA	7.39×10^4	4.4	15 GeV/c single muon filter
$c\bar{c}$	106059	PYTHIA	2.84×10^4	1.5	15 GeV/c single muon filter
J0	105009	PYTHIA	9.75×10^9	0.4	8 < parton p_T < 17 GeV
J1	105010	PYTHIA	6.73×10^8	0.4	17 < parton p_T < 35 GeV
J2	105011	PYTHIA	4.12×10^7	0.4	35 < parton p_T < 70 GeV
J3	105012	PYTHIA	2.19×10^6	0.4	70 < parton p_T < 140 GeV
J4	105013	PYTHIA	8.79×10^4	0.4	140 < parton p_T < 280 GeV
J5	105014	PYTHIA	2.33×10^3	0.4	280 < parton p_T < 560 GeV
J6	105015	PYTHIA	3.39×10^2	0.4	560 < parton p_T < 1120 GeV
J0mu	109276	PYTHIA	9.86×10^9	0.5	8 GeV/c single μ filter $\varepsilon = 7.93 \times 10^{-5}$
J1mu	109277	PYTHIA	6.78×10^8	0.5	8 GeV/c single μ filter $\varepsilon = 1.23 \times 10^{-3}$
J2mu	109278	PYTHIA	4.10×10^7	0.5	8 GeV/c single μ filter $\varepsilon = 5.44 \times 10^{-3}$
J3mu	109279	PYTHIA	2.20×10^6	0.5	8 GeV/c single μ filter $\varepsilon = 1.29 \times 10^{-2}$
J4mu	109280	PYTHIA	8.77×10^4	0.5	8 GeV/c single μ filter $\varepsilon = 2.22 \times 10^{-2}$
J5mu	109281	PYTHIA	2.35×10^3	0.5	8 GeV/c single μ filter $\varepsilon = 2.98 \times 10^{-2}$

Table 2: Monte Carlo samples used in this note. The cross-sections quoted are the ones used to normalize estimates of expected number of events. The cross-sections for the QCD samples ($b\bar{b}$, $c\bar{c}$, and the JX samples) are directly from PYTHIA. Sources for the other cross-sections are discussed in the text.

	Integrated Luminosity (nb^{-1})	
Run number range	W GRL	Z GRL
A-C: 152844-156682	16.65	17.60
D1: 158045-158392	26.89	28.64
D2: 158443-158582	29.03	31.76
D3: 158632-158975	32.85	34.71
D4: 158041-159086	79.40	87.82
D5: 159113	28.04	28.38
D6: 159179-159224	97.05	101.85
Total: 152844-159224	310.0	330.8

Table 1: Integrated luminosity for the runs in periods A to D for the W and Z Good Run Lists. The total integrated luminosity for this dataset is 310.0 nb^{-1} for the W and 330.8 nb^{-1} for the Z.