Разработка базы знаний современного научного эксперимента

Григорьева М.А., к.т.н., Аулов В.А., Голосова М.В., НИЦ «Курчатовский институт»
Губин М.Ю., Томский политехнический университет
Климентов А.А.,к.ф.-м.н., Брукхэвенская национальная лаборатория

Введение

- Особенности DKB
- Источники метаданных
- Архитектура и прототип DKB
- Онтология анализа данных
- Заключение и планы дальнейшего развития

Роль баз знаний в науке

- База знаний (DKB) это система, которая стоит за набором пользовательских интерфейсов и API, агрегируя и соединяя в единое целое метаданные из различных источников, а также улучшает их с помощью добавления к ним связей согласно правилам.
- Одна из главных целей DKB – предоставить научному сообществу **инфраструктуру бузы знаний,** дать учёным простой и быстрый доступ к важной для них информации о экспериментах, облегчить доступ к информации, которая изначально разбросана по различным хранилищам внутри эксперимента.
- DKB должна быть способна *автоматически агрегировать знания* из произвольных, несогласованных, разобщённых ресурсов, включая архивы научных статей, wiki-страницы исследовательских групп, информацию из конференций и совещаний, и *связывать эту информацию* с структурированными техническими данными эксперимента (наборами экспериментальных данных, результатами анализа данных, различными метаданными, описывающими образцы, участвовавшие в эксперименте).
- DKB должно предоставлять учёным согласованное представление жизненного цикла эксперимента.
- Возможные применения DKB:
  - Помощь учёным в настройке экспериментального окружения;
  - Сохранение процесса анализа данных и воспроизведение результатов анализа (например, для учёных вне исходной команды, проводившей эксперимент)
  - Предотвращение удаления данных, легших в основу опубликованных статей, с целью сохранения возможности репликации эксперимента;
  - Обнаружение сходных с уже используемыми научной группой наборов данных, которые с высокой вероятностью содержат информацию, представляющую интерес для учёных;

Источники метаданных [на примере ATLAS]

– Обработка данных:
  - Rucio (Distributed Data Management System)
  - Production System:
    - » DEFT [Database Engine For Tasks]
    - » JEDI [Job Execution and Definition Interface]
  - JIRA ITS (Issue Tracking Service)
  - Репозитории аналитического кода (В ATLAS весь код, используемых для анализа данных, обязательно помещается в системы контроля версий)
  - Google docs (списки наборов данных)
  - Образы виртуальных машин ATLAS (это позволяет сохранить точную конфигурацию ПО)

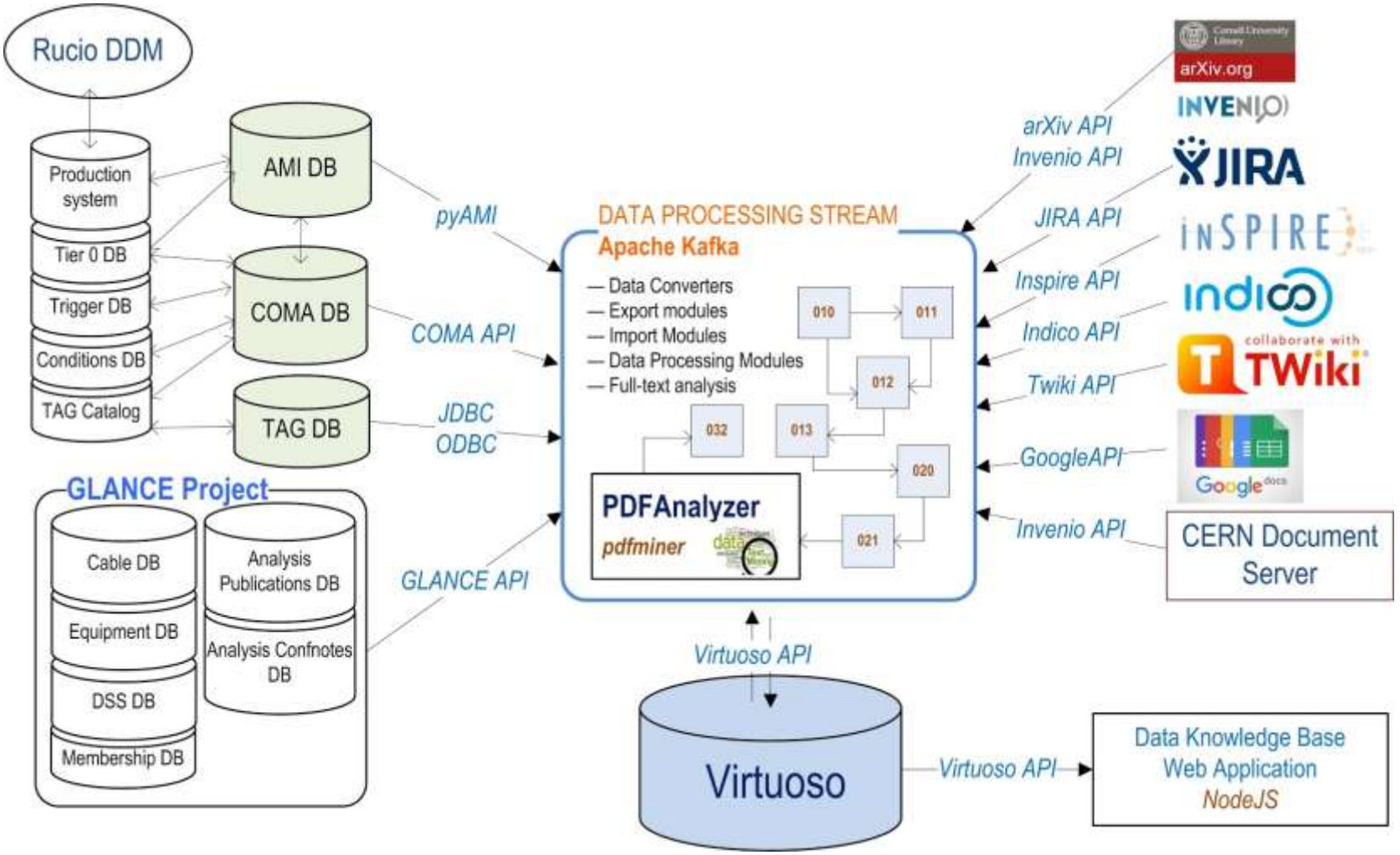
– Научный анализ:
  - Indico (конференции и совещания)
  - CERN Document Server
  - CERN Twiki
  - ATLAS Supporting documents (Internal Notes)

Для их корректной интерпретации, экспериментальные данные должны сопровождаться *вспомогательными метаданными*, которые создаются на каждом шаге обработке экспериментальных данных. Метаданные описывают научные данные и представляют результаты экспериментов, тем самым значительно упрощая публикацию научных данных.

- AMI (Atlas Metadata Interface)
- GLANCE (поисковая система по базам данных ATLAS)

Несмотря на доступность документации, на практике часто оказывается весьма сложной задачей отследить, как именно был получен результат конкретного эксперимента. Необходимая для этого информация часто хранится в нескольких не связанных друг с другом репозиториях, таких как хранилища метаданных, различные репозитории исходного кода, внутренние документы и веб-страницы. Нам же нужно представить весь жизненный цикл анализа данных, начиная с идеи учёного, и заканчивая публикацией.

Архитектура DKB

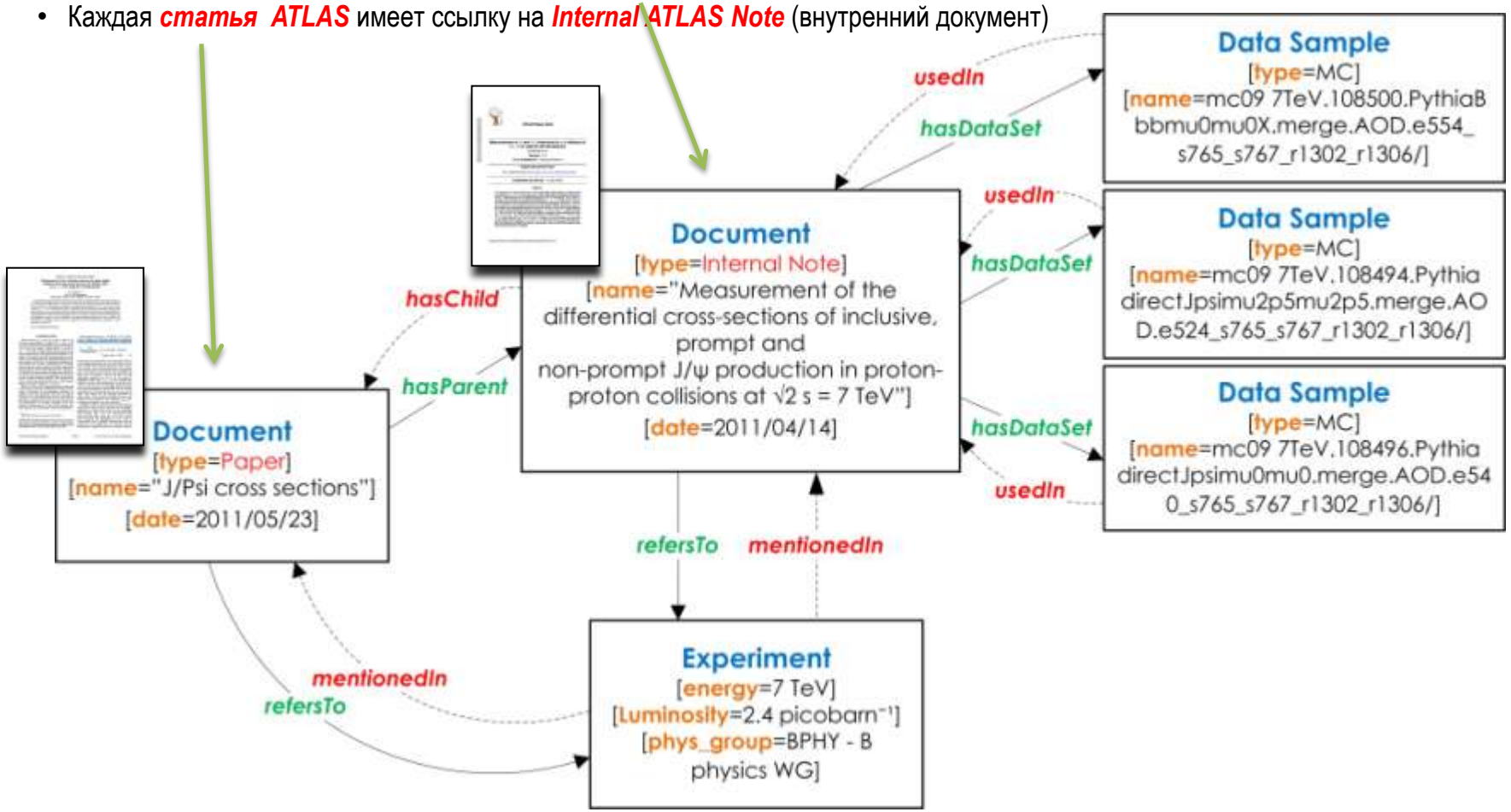


Онтология анализа данных для ATLAS

- Онтология – это словарь терминов и определений предметной области, также она содержит семантические отношения между терминами, позволяя производить логический вывод над сущностями, описанными в онтологии, а также над данными, связанными с содержащимися в ней терминами.
- Онтологический подход к представлению данных позволяет использовать новые подходы к анализу данных, заметно превосходя гибкостью обычные средства поиска по метаданных благодаря гибкости возможностей интеграции информации в онтологическое представление.
- Онтологическое хранилище предоставляет *связанное представление* всех элементов процесса анализа данных в ATLAS.

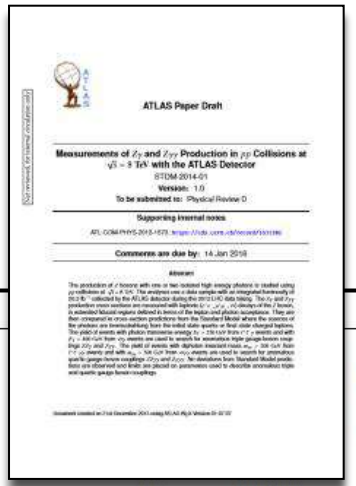
Прототип онтологии эксперимента ATLAS

- Каждая публикация ATLAS основывается на физической гипотезе, которая должна быть подтверждена или опровергнута. Для проверки гипотезы обычно используется пара наборов данных: данные симуляции Монте-Карло и реальные данные с детектора ATLAS. Эти наборы данных обрабатываются цепочкой обработки данных ATLAS, а результаты анализа описываются в статьях и заметках конференций
- Каждая ***статья ATLAS*** имеет ссылку на ***Internal ATLAS Note*** (внутренний документ)



Пример: Представление наборов данных во внутренних документах ATLAS

- Список наборов данных;
- Таблица с атрибутами наборов данных;
- Краткое описание – как были получены наборы данных.


ATLAS Paper Draft

Measurements of $Z\gamma$ and $Z\gamma\gamma$ Production in pp Collisions at $\sqrt{s} = 8$ TeV with the ATLAS Detector

ATDM-2014-01

Version: 1.0

To be submitted to: Physical Review D

Supporting internal notes

Comments are due by: 14 Jan 2016

Sample			σ [pb]	Number of events
mc12_8TeV.126937.PowhegPythia8_AU2CT10_ZZ_4e_mll4_2pt5.merge.NTUP_HSG2.e1280_s1771_s1741_r4829_r4540_p1344			7.68E-02	1099997
mc12_8TeV.126938.PowhegPythia8_AU2CT10_ZZ_2e2mu_mll4_2pt5.merge.NTUP_HSG2.e1280_s1771_s1741_r4829_r4540_p1344			1.76E-01	1599993
mc12_8TeV.126939.PowhegPythia8_AU2CT10_ZZ_2e2tau_mll4_2pt5.merge.NTUP_HSG2.e2372_s1771_s1741_r4829_r4540_p1344			1.75E-01	1099998
mc12_8TeV.126940.PowhegPythia8_AU2CT10_ZZ_4mu_mll4_2pt5.merge.NTUP_HSG2.e1280_s1771_s1741_r4829_r4540_p1344			7.68E-02	1100000
mc12_8TeV.126941.PowhegPythia8_AU2CT10_ZZ_2mu2tau_mll4_2pt5.merge.NTUP_HSG2.e2372_s1771_s1741_r4829_r4540_p1344			1.75E-01	1099998
mc12_8TeV.126942.PowhegPythia8_AU2CT10_ZZ_4tau_mll4_2pt5.merge.NTUP_HSG2.e2372_s1771_s1741_r4829_r4540_p1344			7.69E-02	300000
mc12_8TeV.189591.MCFMPythia8_AU2CT10_gg_ZZ_bkg_4e_m41100_4pt3.merge.NTUP_HSG2.e2761_s1831_s1741_r4829_r4540_p1344			4.94E-04	474998
mc12_8TeV.189592.MCFMPythia8_AU2CT10_gg_ZZ_bkg_2e2mu_m41100_4pt3.merge.NTUP_HSG2.e2761_s1831_s1741_r4829_r4540_p1344	1.0	1.00E+00	9.89E-04	469996
mc12_8TeV.189593.MCFMPythia8_AU2CT10_gg_ZZ_bkg_4mu_m41100_4pt3.merge.NTUP_HSG2.e2761_s1831_s1741_r4829_r4540_p1344	1.0	1.00E+00	4.94E-04	474998
mc12_8TeV.189594.MCFMPythia8_AU2CT10_ggH125p5_ZZ_4e_m41100_4pt3.merge.NTUP_HSG2.e2761_s1831_s1741_r4829_r4540_p1344	1.0	1.00E+00	1.54E-04	464998
mc12_8TeV.189595.MCFMPythia8_AU2CT10_ggH125p5_ZZ_2e2mu_m41100_4pt3.merge.NTUP_HSG2.e2761_s1831_s1741_r4829_r4540_p1344	1.0	1.00E+00	3.09E-04	464397
mc12_8TeV.189596.MCFMPythia8_AU2CT10_ggH125p5_ZZ_4mu_m41100_4pt3.merge.NTUP_HSG2.e2761_s1831_s1741_r4829_r4540_p1344	1.0	1.00E+00	1.54E-04	463997
mc12_8TeV.189597.MCFMPythia8_AU2CT10_ggH125p5_gg_ZZ_4e_m41100_4pt3.merge.NTUP_HSG2.e2761_s1831_s1741_r4829_r4540_p1344	1.0	1.00E+00	6.16E-04	424998
mc12_8TeV.189598.MCFMPythia8_AU2CT10_ggH125p5_gg_ZZ_2e2mu_m41100_4pt3.merge.NTUP_HSG2.e2761_s1831_s1741_r4829_r4540_p1344	1.0	1.00E+00	1.23E-03	459897
mc12_8TeV.189599.MCFMPythia8_AU2CT10_ggH125p5_gg_ZZ_4mu_m41100_4pt3.merge.NTUP_HSG2.e2761_s1831_s1741_r4829_r4540_p1344	1.0	1.00E+00	6.16E-04	460000
mc12_8TeV.185992.MCFMPythia8_AU2CT10_ggZZ_2e2tau_pt3mll4_all_4l.merge.NTUP_HSG2.e3449_s1773_s1776_r4485_r4540_p1344	1.0	1.00E+00	2.59E-04	100000
mc12_8TeV.185993.MCFMPythia8_AU2CT10_ggZZ_2mu2tau_pt3mll4_all_4l.merge.NTUP_HSG2.e3449_s1773_s1776_r4485_r4540_p1344	1.0	1.00E+00	2.59E-04	99999
mc12_8TeV.185994.MCFMPythia8_AU2CT10_ggZZ_4tau_pt3mll4_all_4l.merge.NTUP_HSG2.e3449_s1773_s1776_r4485_r4540_p1344	1.0	1.00E+00	1.52E-05	50000
mc12_8TeV.181767.Pythia8_AU2CTEQ6L1_DPI_Z_Z_4l.merge.NTUP_HSG2.e2538_s1831_s1741_r4829_r4540_p1344	1.0	3.45E-01	2.78E-03	499999
mc12_8TeV.181087.PowhegPythia_P2011C_ttbar_dilepton.merge.NTUP_HSG2.e2091_a188_a205_r4540_p1344	1.2	1.05E-01	2.11E+02	39808972
mc12_8TeV.117650.AlpgenPythia_P2011C_ZeeNp0.merge.NTUP_HSG2.e1477_s1499_s1504_r3658_r3549_p1344	1.0	1.00E+00	8.48E+02	6609984
mc12_8TeV.117651.AlpgenPythia_P2011C_ZeeNp1.merge.NTUP_HSG2.e1477_s1499_s1504_r3658_r3549_p1344	1.0	1.00E+00	2.07E+02	1329498
mc12_8TeV.117652.AlpgenPythia_P2011C_ZeeNp2.merge.NTUP_HSG2.e1477_s1499_s1504_r3658_r3549_p1344	1.0	1.00E+00	6.94E+01	404998
mc12_8TeV.117653.AlpgenPythia_P2011C_ZeeNp3.merge.NTUP_HSG2.e1477_s1499_s1504_r3658_r3549_p1344	1.0	1.00E+00	1.84E+01	109999
mc12_8TeV.117654.AlpgenPythia_P2011C_ZeeNp4.merge.NTUP_HSG2.e1477_s1499_s1504_r3658_r3549_p1344	1.0	1.00E+00	4.64E+00	30000
mc12_8TeV.117655.AlpgenPythia_P2011C_ZeeNp5.merge.NTUP_HSG2.e1477_s1499_s1504_r3658_r3549_p1344	1.0	1.00E+00	1.41E+00	10000
mc12_8TeV.117660.AlpgenPythia_P2011C_ZmumuNp0.merge.NTUP_HSG2.e1477_s1499_s1504_r3658_r3549_p1344	1.0	1.00E+00	8.48E+02	6608490
mc12_8TeV.117661.AlpgenPythia_P2011C_ZmumuNp1.merge.NTUP_HSG2.e1477_s1499_s1504_r3658_r3549_p1344	1.0	1.00E+00	2.07E+02	1334697
mc12_8TeV.117662.AlpgenPythia_P2011C_ZmumuNp2.merge.NTUP_HSG2.e1477_s1499_s1504_r3658_r3549_p1344	1.0	1.00E+00	6.94E+01	404995
mc12_8TeV.117663.AlpgenPythia_P2011C_ZmumuNp3.merge.NTUP_HSG2.e1477_s1499_s1504_r3658_r3549_p1344	1.0	1.00E+00	1.84E+01	110000
mc12_8TeV.117664.AlpgenPythia_P2011C_ZmumuNp4.merge.NTUP_HSG2.e1477_s1499_s1504_r3658_r3549_p1344	1.0	1.00E+00	4.61E+00	30000
mc12_8TeV.117665.AlpgenPythia_P2011C_ZmumuNp5.merge.NTUP_HSG2.e1477_s1499_s1504_r3658_r3549_p1344	1.0	1.00E+00	1.41E+00	10000
mc12_8TeV.117670.AlpgenPythia_P2011C_ZtautauNp0.merge.NTUP_HSG2.e1711_s1581_s1586_r3658_r3549_p1344	1.0	1.00E+00	8.48E+02	6619189
mc12_8TeV.117671.AlpgenPythia_P2011C_ZtautauNp1.merge.NTUP_HSG2.e1711_s1581_s1586_r3658_r3549_p1344	1.0	1.00E+00	2.07E+02	1334898
mc12_8TeV.117672.AlpgenPythia_P2011C_ZtautauNp2.merge.NTUP_HSG2.e1711_s1581_s1586_r3658_r3549_p1344	1.0	1.00E+00	6.92E+01	404795
mc12_8TeV.117673.AlpgenPythia_P2011C_ZtautauNp3.merge.NTUP_HSG2.e1711_s1581_s1586_r3658_r3549_p1344	1.0	1.00E+00	1.83E+01	110000
mc12_8TeV.117674.AlpgenPythia_P2011C_ZtautauNp4.merge.NTUP_HSG2.e1711_s1581_s1586_r3658_r3549_p1344	1.0	1.00E+00	4.66E+00	30000
mc12_8TeV.117675.AlpgenPythia_P2011C_ZtautauNp5.merge.NTUP_HSG2.e1711_s1581_s1586_r3658_r3549_p1344	1.0	1.00E+00	1.39E+00	10000
mc12_8TeV.178354.AlpgenPythia_P2011C_ZeeNp0Excl_Mll10to40_2LeptonFilter5.merge.NTUP_HSG2.e2373_s1581_s1586_r4485_r4540_p1344	1.0	1.00E+00	7.00E+02	5594990
mc12_8TeV.178355.AlpgenPythia_P2011C_ZeeNp1Excl_Mll10to40_2LeptonFilter5.merge.NTUP_HSG2.e2371_s1581_s1586_r4485_r4540_p1344	1.0	1.00E+00	5.11E+01	2299897

Пример: ID наборов данных в внутренних заметках ATLAS

Table 1: Summary of background categories and MC samples used. The cross sections, NLO or NNLO k-factors and filter efficiencies are obtained from the recommended list provided by the SUSY group [54].

Category	Process	MCID	Generator	Remarks
Top	$t\bar{t}$+ boson	119353-55, 119583, 174830-3	MADGRAPH	
	t + boson	179991-2	MADGRAPH	
	tbt	158344	MADGRAPH	
	$t\bar{t}$	110001	MC@NLO	
	Wt	108346	MC@NLO	
	$t\bar{t}+\gamma$	169704-6	MADGRAPH	
WW	WW	126928-36	POWHEG	
	WW via g-g fusion	169471-79	gg2WWJimmy	
	W^*W^*	126988	SHERPA	
	W^*W^*jj	126989	SHERPA	
	WWW^*	167006	MADGRAPH	
ZV	WZ	129477-94	POWHEG	
	ZZ	126937-42, 126949-51, 178411-13	POWHEG	
	ZZ via g-g fusion	116600-3	gg2ZZJimmy	
	$VV \to llqq$	157814-6	SHERPA	
	ZWW^*	167007	MADGRAPH	
	ZZZ	167008	MADGRAPH	
ZX	DY ($m(ll)$ <40 GeV)	178354-68	ALPGEN(Pythia)	
	Z 40 GeV< $m(ll)$ <60 GeV	178369-83	ALPGEN(Pythia)	
	Zcc ($m(ll)$ >60 GeV)	200432-51, 178384-95	ALPGEN(Pythia)	
	Zbb ($m(ll)$ >60 GeV)	200332-51, 178396-407	ALPGEN(Pythia)	
	Z ($m(ll)$ >60 GeV)	147105-10, 147113-18, 147121-6	ALPGEN(Pythia)	
	$Z\gamma\gamma$	145161-2	SHERPA	
Higgs	g-g fusion $H \to 2l4l$	160155, 160655, 161005	POWHEG	
	g-g fusion $H \to \tau\tau$	161555-66-77	POWHEG	
	VBF	160205, 160705, 161055	POWHEG	
	$WH \to 2l4l$	160255, 160755, 161105	Pythia8	
	$WH \to VV$	160505, 161805	Pythia8	
	$ZH \to 2l4l$	160305, 160805, 161155, 167418	Pythia8	
	$ZH \to ll + xx$	160555, 161675-86-97	Pythia8	
	$t\bar{t}H$	161305, 161708-19-30, 169072	Pythia8	

Table 2: The list of $H \to \tau^{\pm}\mu^{\mp}$ and $H \to \tau^{\pm}e^{\mp}$ simulated samples.

Production	Decay	MCID
ggF	$\tau\mu$	189890
VBF	$\tau\mu$	189889
VH	$\tau\mu$	189891-2
$t\bar{t}H$	$\tau\mu$	-
ggF	τe	189894
VBF	τe	189893
VH	τe	189895-6
$t\bar{t}H$	τe	-

ID наборов данных симуляций Монте-Карло

ProductionSystem: база данных DEFT

Метаданные, которые содержатся в ATLAS Internal Notes:

- LHC Energy Run
- Светимость LHC
- Год, Run Number, Периоды
- Тип столкновений (p-p, Pb-Pb)
- Генераторы Монте-Карло
- Используемые триггеры
- Статистика
- Наборы данных
  - Экспериментальные данные
  - Данные симуляций Монте-Карло
  - Сигнал
  - Фон
- Версия ПО
- Условия эксперимента

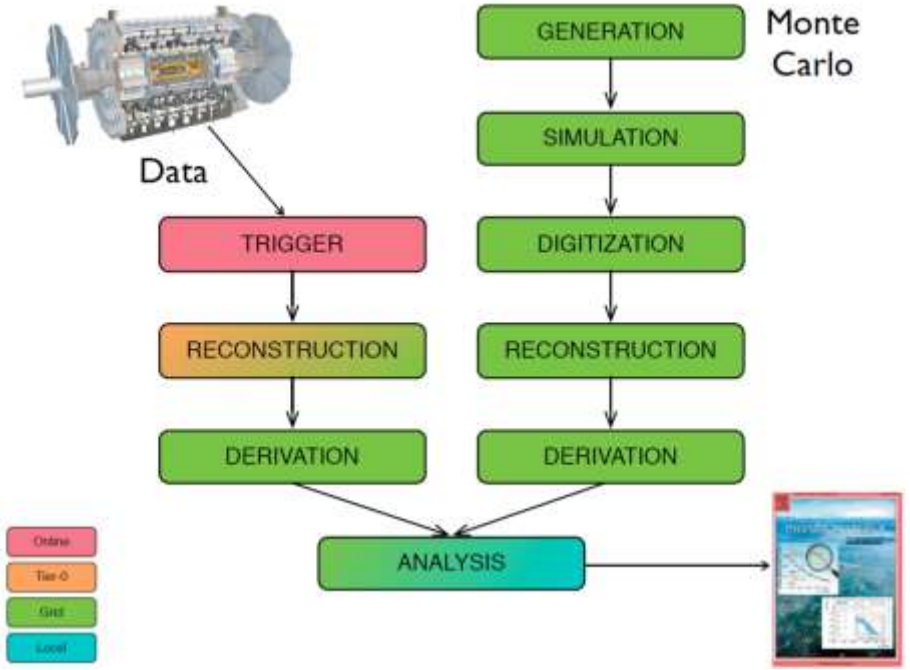
Доступны в метаданных статей

Доступны только в полном тексте документов

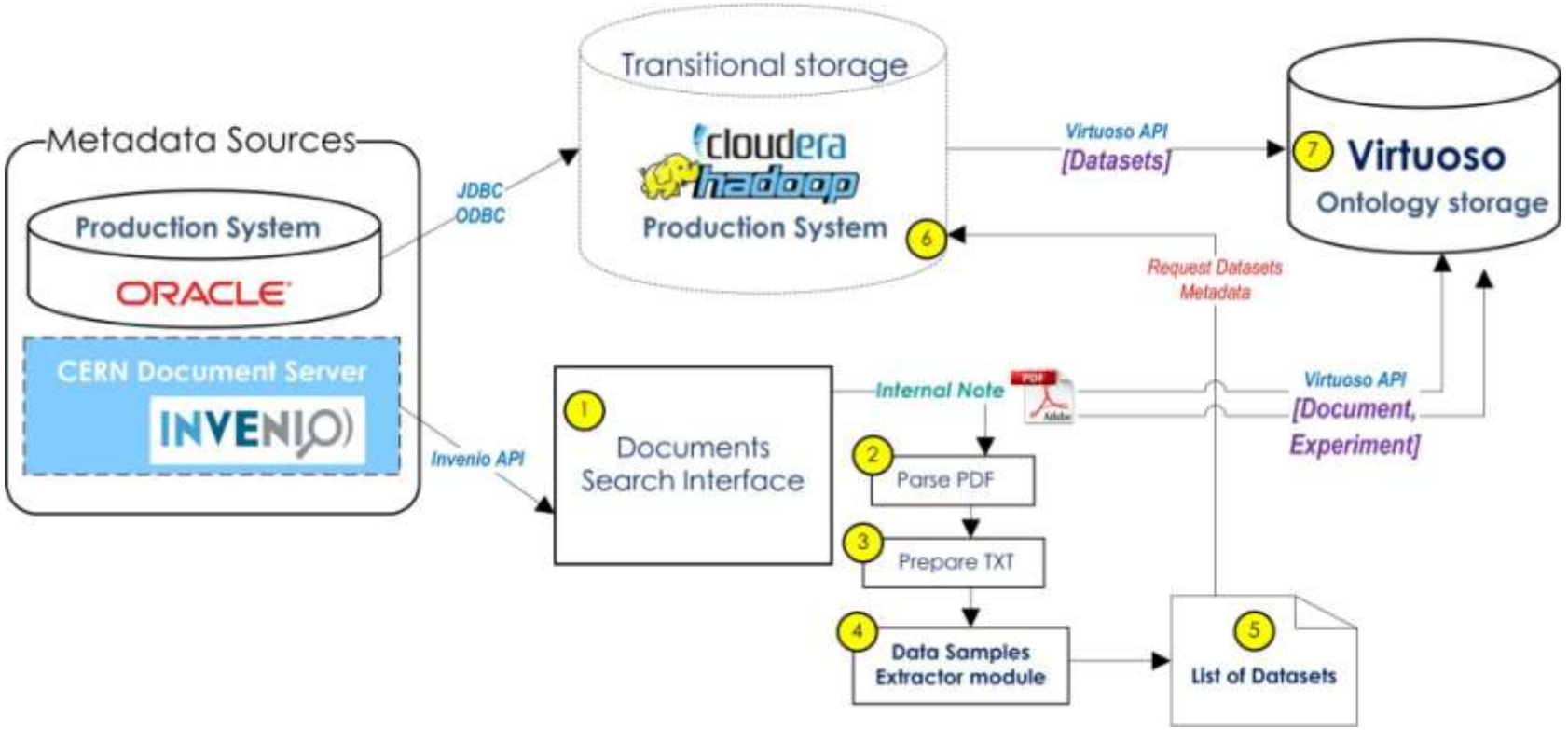
Метаданные эксперимента должны автоматически извлекаться из содержимого научных статей и внутренних документов ATLAS.

Формализация описания анализа данных

В целом, описания анализа данных во внутренних заметках ATLAS хорошо структурированы. Авторы используют очень точные предложения, слова и фразы для описания, как и на каких данных проводился эксперимент.
Это позволяет нам аннотировать текст, получив из него нужную информацию.



Процесс извлечения метаданных для наборов данных



1. Параметрический поиск статей и внутренних документов ATLAS в CDS
2-4. Анализ полнотекстового документа, вставка результатов анализа в базу знаний.
5. Результат – список наборов данных
6. Запрос к хранилищу Hadoop чтобы получить метаданные для наборов данных
7. Вставка метаданных наборов данных в хранилище Virtuoso

Заключение

- Разработка прототипа хранилища метаданных физического эксперимента (на примереATLAS):
  - Онтология:
    - Разработан прототип онтологии ATLAS Data Analysis с основными классами: документ, набор данных, сотрудник, эксперимент
    - Развёрнуто хранилище Virtuoso на базе ТПУ
  - Промежуточное хранилище Hadoop развёрнуто в Курчатовском институте
    - Метаданные Production System экспортированы из БД Oracle в Hadoop
  - Внутренние документы:
    - Разработаны инструменты для извлечения метаданных из полнотекстовых документов
    - Разработан модуль извлечения наборов данных из внутренних заметок
  - В разработке:
    - Веб-интерфейс
    - Инструменты для работы с данными онтологии через Virtuoso API
    - Поисковый интерфейс для документов ATLAS с использованием Invenio API

План дальнейшего развития

Разработать действующий прототип архитектуры DKB:

- Хранилище онтологий (на базе Virtuoso) с онтологией “Document-Dataset-Experiment”;
- Простой веб-интерфейс, который позволит пользователям искать метаданные публикаций, экспериментов и наборов данных;
- Инструментарий для извлечения информациии из внутренних документов;
- Инструменты для обработки данных в Virtuoso.

Информация о финансовой поддержке

Эта работа была выполнена при поддержке госзадания «Наука» №3997