Разработка прототипа базы знаний современного научного сообщества

М. А. Григорьева[1,а], В. А. Аулов[1,б], М. В. Голосова[1,в], М. Ю. Губин[2,г], А. А. Климентов[3,д]

[1] Национальный исследовательский центр «Курчатовский институт», Москва, пл. Академика Курчатова, д.1

[2] Национальный исследовательский Томский Политехнический Университет, Томск, проспект Ленина, д. 30

[3]Брукхэйвенская Национальная Лаборатория, США, Брукхэйвен авеню, Аптон, NY 11973

E-mail: [а] maria.grigorieva@cern.ch, [б] vasiliyaulov@gmail.com, [в] marina.golosova@cern.ch, [г] maksim.gubin@cern.ch, [д] alexei.klimentov@cern.ch

Современные научные эксперименты с интенсивной обработкой данных имеют длительный жизненный цикл, сложную распределенную программно-аппаратную инфраструктуру, в которой хранятся данные сотни петабайт и обратываются экзабайты данных. Все стадии жизненного цикла эксперимента сопровождаются вспомогательными метаданными, необходимыми для мониторирования процессов обработки и управления, а также для воспроизводимости результатов эксперимента. В большинстве научных сообществ метаданные, описывающие цепочки анализа и обработки данных, и метаданные о публикации научных результатов, существуют независимо друг от друга. Кроме того, чтобы воспроизвести или подтвердить результаты уже проведенного эксперимента, ученым бывает необходимо провести исследования при тех же условиях, проверить результаты обработки наборов данных новой версией программного обеспечения, или опробовать новые алгоритмы. Вот почему вся информация об анализе данных должна быть сохранена, начиная от выдвигаемой гипотезы и цепочки преобразования данных, и до публикации результатов. Описанная в работе база научных знаний (Data Knowledge Base - DKB) обеспечивает хранение и быстрый доступ к релевантной научной и вспомогательной метаинформации. В основе DKB лежит онтология научных исследований в области физики высоких энергий. Архитектура DKB имеет два уровня хранения данных: хранилище Hadoop, в котором данные от различных источников метаданных интегрируются, агрегируются и обрабатываются, и онтологическое хранилище Virtuoso, в котором сохраняются все извлеченные данные. Агенты DKB обрабатывают и агрегируют метаданные из систем управления и обработки данных, интерфейсов поиска метаданных, архивов тезисов конференций и статей. Дополнительно эти метаданные связываются с соответствующими интернет-ресурсами (в системах коллективного аннотирования и документирования Twiki, редактирования документов и таблиц – Google Docs), и информацией, извлекаемой из полных текстов научно-исследовательской документации. DKB агенты позволяют извлекать, агрегировать и интегрировать всю необходимую метаинформацию автоматически, избавляя ученых от необходимости подробно аннотировать каждый компонент эксперимента.

Ключевые слова: база научных знаний, онтология, RDF-хранилище, Virtuoso, жизненный цикл эксперимента, цепочка обработки данных, научная публикация.

Работа выполнена при финансовой поддержке гранта Правительства Российской Федерации, выделенного на конкурсной основе для государственной поддержки научных исследований, проводимых под руководством ведущих ученых в российских образовательных учреждениях высшего профессионального образования (постановление правительства № 220 от 9 апреля 2010 года), по договору № 14.Z50.31.0024.

1. Введение

Одной из основных проблем развития современной науки стало стремительное нарастание объемов информации с экспериментальных установок, метаинформации, версий программного обеспечения, используемого для анализа, обработки и хранения данных. Особенно актуальна эта проблема для научных исследований, проводимых и планируемых на базе крупнейших установок, таких как NICA, XFEL, LHC, ITER, FAIR и др. Научные коллаборации, ведущие исследования на таких установках, включают ученых из десятков стран, а сама программа исследований продолжается 15-20 лет. Сотни петабайт данных, генерируемых с помощью научных установок, требуют распределенной системы управления, хранения и доступа, а также распределенных центров высокопроизводительных вычислений для анализа и обработки данных.

Учитывая длительность и сложность современных исследований, каждый этап проведения научного эксперимента, от формулирования гипотезы и выбора методов исследования, проведения эксперимента в заданных условиях аппаратного, программного и физического окружения, до обсуждения результатов на совещаниях и конференциях, и публикации результатов, сопровождается сбором и хранением большого количества вспомогательной информации (метаданных), регистрируемой в различных репозиториях. Однако эти репозитории существуют независимо друг от друга и практически не имеют семантической связи, что затрудняет автоматизацию сопровождения эксперимента.

Работа, представленная в данной статье, посвящена разработке базы научных знаний (Data Knowledge Base - DKB) - платформы, позволяющей интегрировать метаданные из структурированных и документальных источников, и обеспечивающей удобную инфраструктуру хранения и доступа к метаданным научного исследования. DKB позволит интегрировать в единое информационное пространство метаинформацию всех стадий жизненного цикла научных исследований. Данная статья ориентирована на источники метаданных эксперимента АТЛАС на БАК, но возможное применение такой базы знаний может быть в любом из современных экспериментов, специфика АТЛАС важна только на этапе как происходит агрегирование исходной информации. В статье будут представлены: основные источники метаданных эксперимента ATLAS, метод формализации предметной области – онтологическая модель научного исследования, прототип архитектуры базы научных знаний, метод извлечения метаданных из полных текстов научных документов и статей, и предложена технология автоматизации рабочего потока экспорта и импорта данных в RDF-хранилище.

2. Метаданные эксперимента ATLAS

Для формирования общей картины научного исследования, а также всей вспомогательной метаинформации, было проведено исследование источников метаданных в эксперименте ATLAS. Условно, их можно разделить на две группы:

1) ***метаданные процесса распределенной обработки и анализа данных:***

- Rucio (Distributed Data Management System) – распределенная система управления и передачи данными, обеспечивающая формирование наборов данных и управление передачей информации в распределенной компьютерной среде;
- Production System – система обработки и анализа данных и управления загрузкой, распределяющая задания обработки и анализа в гетерогенной среде. Она состоит из трех компонент: DEFT (Database Engine For Tasks) – СУБД управления заданиями, JEDI (Job Execution and Definition Interface) – СУБД управления задачами, PanDA (Production and Distributed Analisys System) – система управления рабочим потоком;
- JIRA ITS (Issue Tracking Service) – сервис отслеживания ошибок;

– Исходные коды анализа и обработки данных хранятся в репозиториях системы контроля версий;

– Исследовательские группы хранят списки наборов экспериментальных и моделированных данных в Google Docs и в Twiki;

– Программные и аппаратные конфигурации хранятся в образах виртуальных машин ATLAS;

2) метаданные о процессе представления и публикации научных результатов:

Эта группа метаданных в основном формируется из документальных источников: препринты, статьи, труды конференций и пр., хранящиеся в системе документооборота ЦЕРН (CERN Document Server), InSpire, на страницах Twiki и JIRA, а также в системе Indico.

Все эти системы независимы и информация между ними не синхронизирована. Частичную связность между ними обеспечивают системы AMI[1] (Metadata Interface and database) – фреймворк для каталогизации и поиска метаданных ATLAS, GLANCE[2] - мощный поисковый движок для коллаборации ATLAS, агрегирующий метаинформацию из различных источников. Однако и они не позволяют представить весь жизненный цикл научного исследования от физической гипотезы до анализа результатов и научной публикации [Григорьева, Голосова, …, 2015].

3. Прототип архитектуры базы научных знаний

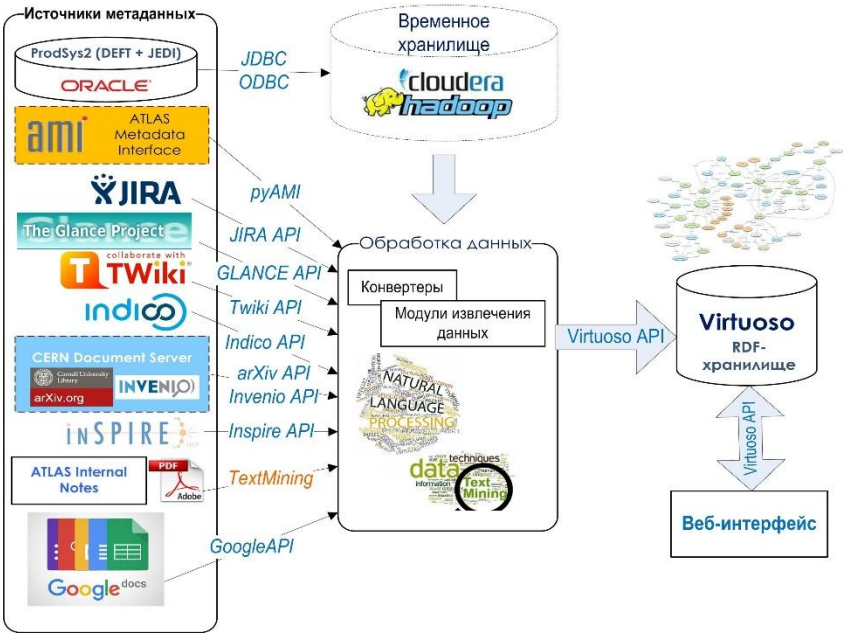

Рисунок 1. Прототип архитектуры DKB

В общем виде, база научных знаний представляет собой группу сервисов обработки данных из структурированных и документальных источников, и их интеграции в онтологическом хранилище. Прототип архитектуры разрабатываемой базы научных знаний приведен на рисунке 1. Онтологический подход, благодаря целостному описанию предметной области, позволяет значительно расширить понимание взаимосвязей между различными этапами, вспомогательными подсистемами и документами научного эксперимента, и, возможно, найти связи, которые ранее были недоступны, не замечены или неизвестны [Allemang, Hendler, 2011].

[1] http://ami.in2p3.fr/index.php/en/

[2] https://atglance.web.cern.ch/atglance/

4. Разработка онтологической модели научного исследования

Чтобы построить полную онтологическую модель научного исследования, необходимо формализовать описание всех имеющихся в исследовании этапов, параметров, процессов, участников и других сущностей. На сегодняшний день существуют онтологии, позволяющие описывать как научные публикации (Dublin Core, SKOS, CERIF, и др), так и научные эксперименты в целом (EXPO, Detector Final State, CSMD) [Soldatova, King, 2006]. Учитывая наработки существующих онтологических моделей, и специфику метаданных эксперимента ATLAS, был разработан фрагмент онтологического представления научного исследования, который позволяет параметрически описывать эксперименты, документы, включая их наследование, авторов, и наборы научных данных.

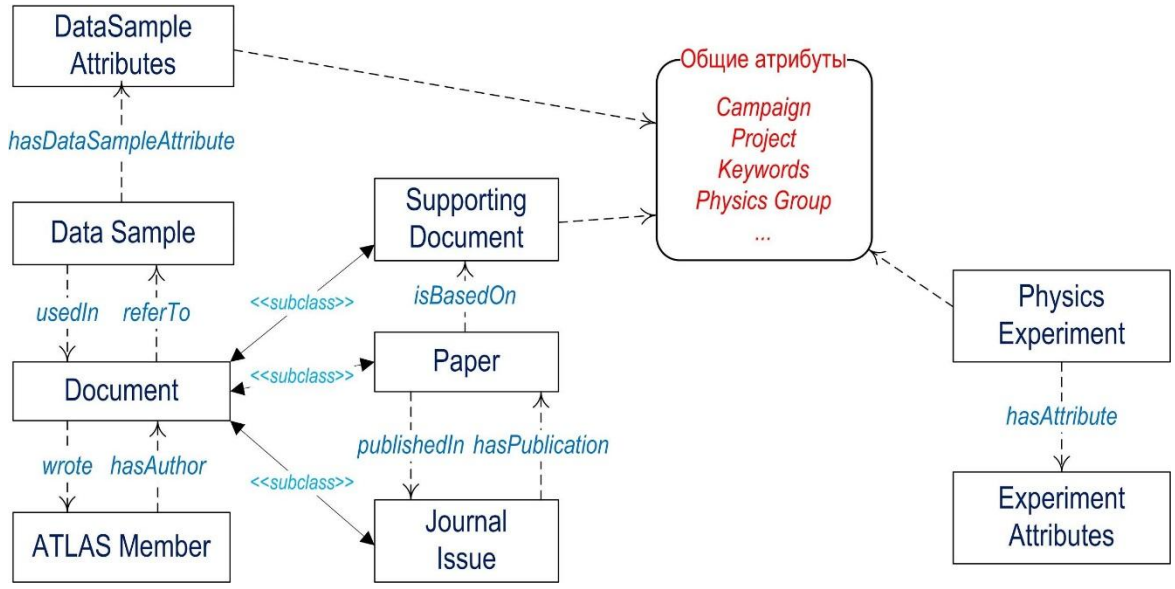

Рисунок 2. Фрагмент онтологии эксперимента ATLAS

Основной сущностью онтологической модели является документ (Document). Документы могут быть нескольких типов: публикации, внутренняя документация, выпуск журнала, и др. Все публикации опубликованы в соответствующем выпуске научных журналов (JournalIssue) - связь “publishedIn”/”hasPublication”, и основаны на соответствующих внутренних документах (связь “isBasedOn”). Любой документ имеет автора (ATLASMember) - связь “hasAuthor”/”wrote”. Внутренние документы имеют ссылки на сэмплы данных (Data Sample) - связь “referTo”/“useIn”. Документы и сэмплы данных могут быть отнесены к определенному физическому эксперименту (PhysicsExperiment) по набору общих атрибутов. В настоящее время такими атрибутами определены: название проекта физического анализа (Project), и кампании (Campaign), ключевые слова (Keywords), название физической группы (Physics Group). В дальнейшем набор общих атрибутов будет увеличиваться, что позволит обеспечить более сильную связность между метаданными.

5. Извлечение информации из текстов научных публикаций

Исходные тексты научных документов находятся в формате PDF. Данный формат весьма удобен для восприятия человеком, но, к сожалению, мало приспособлен для машинной обработки. После изучения имеющихся инструментов по работе с PDF, в качестве средства первич-

ной обработки был выбран PDFMiner. Он позволяет извлекать текст из PDF и сохранять его в одном из нескольких форматов, среди которых используются TXT и XML.

В формате TXT все имеющиеся в PDF символы (с добавлением пробелов, пустых строк и т.д.) конвертируются в сплошной поток текста, независимо от размера шрифта. В результирующем тексте довольно просто (по сравнению с другими подобными программами) отделить куски основного текста документа от разнообразного «мусора», такого как номера строк. Именно это послужило причиной выбора PDFMiner в качестве средства первичной обработки.

В формате XML перечисляется каждый символ документа, с указанием положения на странице и размера шрифта. Данный формат сложнее анализировать, но содержащаяся в нем дополнительная информация необходима в тех случаях, когда нужно учитывать относительное положение символов (например, в таблицах).

Для дальнейшей обработки был разработан анализатор - PDFAnalyzer, который осуществляет поиск полезной информации в извлеченном тексте. Простейший случай - поиск датасетов (то есть наборов данных) в основном тексте статьи с помощью регулярных выражений, возможный благодаря существованию строгой номенклатуры именования датасетов. В некоторых документах датасеты содержатся в таблицах, обрабатывать которые значительно сложнее - необходимо проанализировать каждую страницу с таблицей в формате XML, отделить данные таблицы от всего остального, а затем сконструировать из них столбцы и строки.

Анализ текстов сильно затрудняется прежде всего двумя факторами: упомянутой вначале сложностью конвертации PDF (к примеру, при мелком размере шрифта символы подчеркивания могут считываться как пробелы) и тем, что документы писались с расчетом на чтение людьми, поэтому авторы документов могут именовать разделы с данными и колонки таблиц в довольно свободной форме, употреблять интервалы в названиях датасетов и т.д. Тем не менее, разработанный анализатор обрабатывает 70% имеющихся PDF документов.

6. Реализация и автоматизация рабочего потока при экспорте и импорте данных в RDF-хранилище Virtuoso

Задачу заполнения базы научных знаний и поддержания её в актуальном состоянии можно разбить на три подзадачи:

- получение метаданных из внешних источников знаний;
- обработка полученных метаданных (выделение нужной информации и приведение её к виду, в котором она будет храниться в конечных хранилищах базы научных знаний);
- сохранение обработанных метаданных в конечных хранилищах.

Обработка данных может состоять из нескольких последовательных действий. В простейшем случае процесс можно представить в следующем виде:

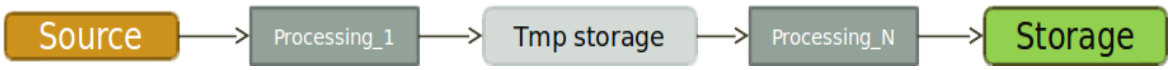

Рисунок 3. Схематичное представление потока данных от источника к конечному хранилищу

Однако в реальности на каком-то этапе обработки исходных метаданных могут потребоваться метаданные из нескольких внешних источников, а результаты обработки - быть задействованы в нескольких последующих шагах обработки. Как правило, основная сложность заключается не в том, чтобы реализовать каждый шаг-разветвлённого потока данных, а в том, чтобы связать их все воедино, гарантировать своевременность выполнения каждого шага, а также то, что все входные данные будут обработаны (и не более одного раза) и помещены в конечные хранилища.

После того, как было реализовано несколько ключевых процессов для потока данных из источников метаинформации ATLAS, схема потока данных приняла следующий вид:

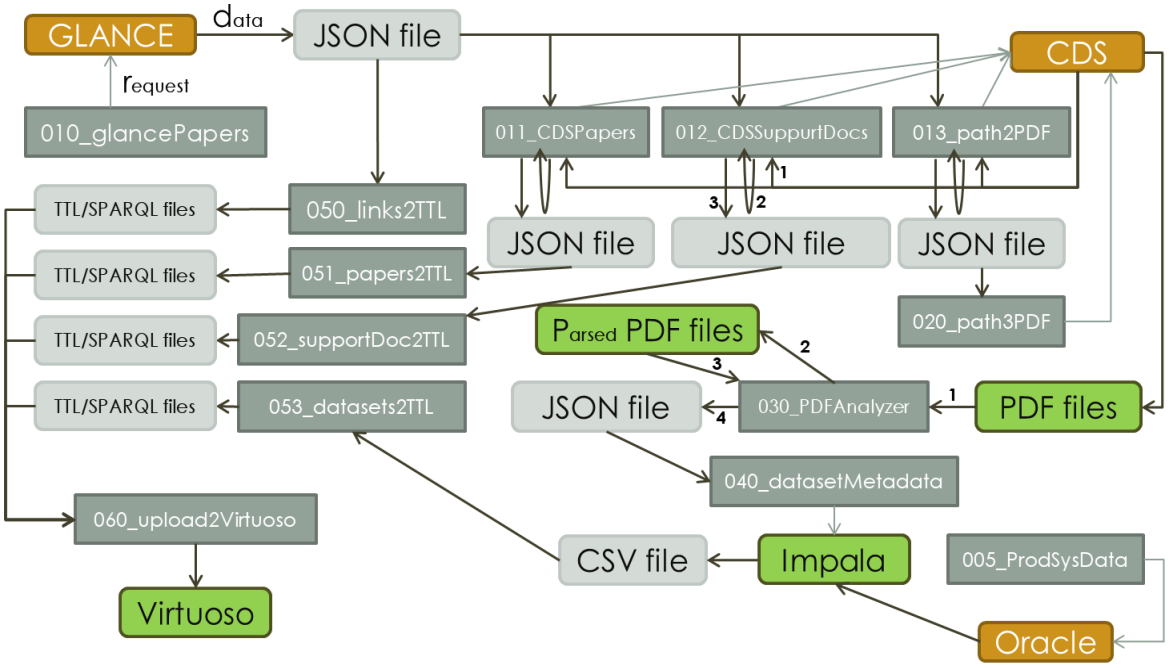

Рисунок 4. Схема потока данных для прототипа базы научных знаний эксперимента ATLAS

Даже для прототипа, работающего с небольшим количеством источников данных, автоматизация потока данных стала насущной необходимостью.

Для решения такого рода задач существуют специальное программное обеспечение - менеджеры потоков данных, или системы обмена сообщениями между компонентами программной системы (RabbitMQ, Apache Flume, Apache Kafka, и др.). Сравнительный анализ нескольких такие систем показал, что для организации потока данных, который включает в себя не только передачу, но и обработку данных, необходима также и система потоковой обработки данных (Apache Spark или Apache Storm).

Одна из изученных систем, Apache Kafka [Kreps, Narkhede, Rao, 2011], с весны 2016 года предоставляет оба функционала - как передачи, так и потоковой обработки данных. Именно это определило выбор технологии для реализации управления потоками метаданных от источников к конечным хранилищам для DKB.

Было разработано приложение, использующее библиотеку Kafka Streams (компонент Apache Kafka, реализующий потоковую обработку данных), позволяющее встроить уже созданные программные модули в архитектуру Kafka (независимо от того, на каких языках программирования модули были написаны). Помимо существенного ускорения процесса автоматизации управления потоком данных за счёт использования уже существующих наработок, это также существенно расширяет возможности разработчиков по созданию новых программных модулей, не ограничивая их использованием одного языка программирования.

7. Заключение

Разработка системы научных знаний является критически важной для развития и дальнейшей деятельности крупных научных коллабораций. До последнего времени проблема автоматизированного сопровождения эксперимента и воспроизводимости результатов решалась через разработку набора сервисов, позволяющих связать между собой различные источники метаданных. Однако, в условиях стремительного роста объемов метаданных, очевидна необходимость формализованного описания всех метаданных и хранения их в онтологическом храни-

лище. Дальнейшее развитие DKB будет связано с расширением онтологической модели, подключением новых источников метаданных, разработкой сервисов и рабочих потоков обработки данных, и усовершенствованием механизмов анализа PDF документов.

Данная работа выполнена при поддержке гранта Правительства Российской Федерации (постановление правительства № 220 от 9 апреля 2010 года), договор № 14.Z50.31.0024.

Список литературы

Григорьева М., Голосова М., Рябинкин Е., Климентов А. Экзабайтное хранилище научных данных // Открытые системы. СУБД. – 2015. – Т. 4. – С. 14-17.

Grigorieva M., Golosova M., Ryabinkin E., Klimentov A. Ekzabajtnoe hranilishche nauchnyh dannyh // Otkrytye sistemy. SUBD. – 2015. – Vol. 4. – P. 14-17.

Soldatova L., King R. An Ontology of Scientific Experiments // Journal of the Royal Society Interface. – 2006. – Vol. 3, Issue 11. – P. 795-804.

Allemang D., Hendler J. Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL // Elsevier. — 2011. ISBN 978-0-12-385965-5.

Kreps J., Narkhede N., Rao J. «Kafka: A Distributed Messaging System for Log Processing» // NetDB workshop. – 2011. URL: http://research.microsoft.com/en-us/um/people/srikanth/netdb11/netdb11papers/netdb11-final12.pdf.

Data Knowledge Base Prototype for Modern Scientific Collaborations

M. A. Grigorieva[1,a], V. A. Aulov[1,b], M. V. Golosova[1,c], M. Y. Gubin[2,d], A. A. Klimentov[3,e]

[1] National research Center "Kurchatov Institute", 1, Akademika Kurchatova sq., Moscow, Russia

[2] National research Tomsk Polytechnic University, 30, Lenin Avenue, Tomsk, Russia

[3] Brookhaven National Laboratory, Upton, NY 11973 5000, USA

E-mail: [a] maria.grigorieva@cern.ch, [b] vasiliyaulov@gmail.com, [c] marina.golosova@cern.ch, [d] maksim.gubin@cern.ch, [e] alexei.klimentov@cern.ch

The most common characteristics of large-scale modern scientific experiments are long lifetime, complex experimental infrastructure, sophisticated data analysis and processing tools, peta- and exascale data volume. All stages of an experiment life cycle are accompanied with the auxiliary metadata required for monitoring, control and scientific results replicability and reproducibility. The actual issue for the majority of scientific communities is a very loose coupling between metadata describing data processing cycle, and metadata representing annotations, indexing and publication of the experimental results. Besides, to reproduce and to verify some previous data analysis, it's very important for the scientists to conduct studies under the same conditions or to process data collection with new software releases or/and algorithms. That's why all information about data analysis process must be preserved, starting from the initial hypothesis following by processing chain description, data collection, initial results presentation and final publication. A knowledge-based infrastructure (Data Knowledge Base - DKB) gives such possibility and provides fast access to relevant scientific and accompanying information. DKB is functioning on the basis of HEP data analysis ontology. The architecture has two data storage layers: Hadoop storage, where data from many metadata sources are integrated and processed to obtain knowledge-based characteristics of all stages of the experiment, and Virtuoso RDF-storage, where all extracted data are registered. DKB agents process and aggregate metadata from data management and data processing systems, metadata interface, conference notes archives, workshops and meetings agendas, and publications. Additionally, this data is linking with the scientific topic documentation pages (such as Twiki pages, Google documents, etc) and information extracted from full texts of experiment supporting documentation. In this way, rather than require the physicists to annotate all meta information in details, DKB agents will extract, aggregate and integrate all necessary metadata automatically.

Keywords: data knowledge base, ontology, RDF-storage, Virtuoso, experiment life-cycle, data processing chain, scientific publication.

The work was supported by the Russian Ministry of Science and Education under contract No. 14.Z50.31.0024.

-