АНТОПОЛЬСКИЙ Александр Борисович *- доктор технических наук, главный научный сотрудник Института научной и педагогической информации РАО, член редколлегии журнала «Информационные ресурсы России»*
Адрес: 142432, Московская обл., Ногинский р-он, пос. Черноголовка, Школьный б-р, 1
e-mail: ale5695@yandex.ru

АТАЕВА Ольга Муратовна *- младший научный сотрудник Вычислительного центра РАН (ВЦ РАН)*
Адрес: 119333, г. Москва, ул. Вавилова, 40
e-mail: oli@ultimeta.ru

СЕРЕБРЯКОВ Владимир Алексеевич *- доктор физико-математических наук, профессор, заведующий отделом систем математического обеспечения ВЦ РАН*
Адрес: 119333, г. Москва, ул. Вавилова, 40
e-mail: serebr@ccas.ru

Среда интеграции данных научных библиотек, архивов и музеев «LibMeta»*

Введение

Одной из важных современных тенденций развития информационного пространства является тенденция на интеграцию разнородных ресурсов в рамках распределенных информационных систем, в том числе, распределенных электронных библиотек. Эта тенденция особенно заметна в информационной среде науки и культуры. Главная причина этой тенденции заключается в том, что основные владельцы научной и культурной информации, а это, прежде всего, библиотеки, музеи и архивы (иногда их также называют «институты памяти»), находятся в сфере информатизации под влиянием противоречивых стимулов.

С одной стороны, эти учреждения активно заинтересованы в выходе в интернет, представлении своей информации в цифровой форме, в увеличении доступа пользователей к своим ресурсам и вообще в расширении спектра электронных услуг. Эта тенденция поддерживается государством: в рамках программы «электронного правительства» для государственных и муниципальных институтов памяти установлен обязательный набор «электронных услуг», который должен постоянно расширяться. В результате число электронных коллекций по научному и культурному достоянию, создаваемых институтами памяти, быстро растет и уже исчисляется тысячами.

С другой стороны, подавляющее большинство научных учреждений не обладает финансовыми, технологическими и особенно кадровыми ресурсами для создания современных высокоразвитых электронных библиотек. При этом попытки государственного централизованного управления созданием электронных коллекций в институтах памяти терпят провалы. Достаточно вспомнить историю с электронным каталогом Государственного музейного фонда, работу по которому не могут организовать уже 12 лет.

С третьей стороны, в институтах памяти весьма развито психологически оправданное стремление сохранить контроль над цифровыми ресурсами, создаваемыми на основе фондов этих институтов. Это касается правовых и коммерческих аспектов, но в наибольшей степени определяется соображениями престижа и продвижения собственного бренда.

Отвечая этим тенденциям, специалисты по информационным технологиям активно развивают распределенные информационные системы. Основным архитектурным принципом этих систем является децентрализованное создание и хранение цифровых данных в сочетании с централизованной базой метаданных, а также общими сервисами навигации и поиска во всем распределенном информационном пространстве. При этом метаданные могут и должны также храниться на сервере участника, обеспечивая более специфические возможности поиска в локальной системе,

Наиболее известной в мировой практике распределенной информационной системой по научному и культурному наследию является Европеана [1]. Другим примером такого рода могут служить распределенные научные социальные сети, создаваемые в рамках Инициативы открытых архивов, например, Соционет [2].

Наиболее развитой и универсальной информационно-технологической средой для проектов интеграции электронных коллекций научной и культурной информации, прежде всего создаваемых в академической среде, является портал «LibMeta», созданный коллективом специалистов Вычислительного центра РАН.

* *Настоящая работа поддерживается грантами РФФИ № 11-07-00331-а и № 11-07-00286-а.*

Основные возможности LibMeta

Портал «Libmeta» - это стандартизированная и децентрализованная среда управления информацией электронных фондов, разработанная для интеграции ресурсов институтов памяти (библиотек, архивов и музеев), содержащих объекты научного наследия и связанных с ними метаданными из различных источников, облегчающая обмен информацией и ее совместное использование посредством интернета.

Такой подход к управлению информацией электронных фондов институтов памяти имеет целью предоставить широкому сообществу пользователей средства для простого и своевременного доступа к имеющимся данным, обеспечить навигацию пользователя и сквозной поиск объектов научного и, шире, культурного наследия в фондах различных институтов памяти [3].

Благодаря тому, что портал «LibMeta» построен на базе ИС «НИ РАН» [4], являющейся базовым инфраструктурным компонентом ЕНИП [5], он может интегрироваться в ЕНИП с предоставлением расширения схемы метаданными научного наследия институтов памяти. ЕНИП разрабатывается в рамках программы создания и объединения информационных систем подразделений и научных институтов РАН для удовлетворения потребностей научных сотрудников как в части поиска информации, так и в представлении собственной информации в сети Интернет.

Основной профиль метаданных ЕНИП включает в себя большое количество различных схем описания:

- базовая схема описания персоны;
- базовая схема организационных единиц;
- общая схема описания должностей;
- базовая схема проектов;
- базовая схема описания библиографической информации;
- схема мероприятий.

Помимо базовых схем, существуют также и расширенные, в которых производится уточнение введенных классов и их атрибутов.

В контексте этой работы наибольшую значимость представляют схемы описания персон, публикаций и организаций.

Основные типы ресурсов

Система поддерживает следующие основные типы ресурсов: Предметы (музейные), Единица описания (архивы) и дополнительные типы ресурсов, такие как Организация, Персона, Публикация, Проект и различные рубрикаторы и классификаторы. Требование стандартизации метаданных физических музейных предметов и их мультимедийных (фото-, видео-, аудио-) представлений привело к созданию дополнительных прикладных профилей поддержки музейной деятельности.

В отличие от публикаций описания музейных объектов могут значительно отличаться в различных музеях, и здесь невозможно обеспечить всеобъемлющий набор необходимых свойств. В связи с этим реализуется возможность определения дополнительных свойств в виде связей с двумя вспомогательными объектами: Дополнительные свойства и Значения дополнительных свойств. Также добавлен класс Медиа-объект, предназначенный для описания медиа-объекта как единого целого, состоящего из частей данных с различной функциональной нагрузкой, и класс Часть медиа-объекта, позволяющий в пределах одного целого медиа-объекта иметь несколько частей с различной функциональной нагрузкой.

В LibMeta также поддерживаются Коллекции, которые позволяют хранить классические ресурсы (архивные, музейные) и иметь любые вложенные наборы объектов (выставочные, выездные, по хранению и пр.).

Единица описания в контексте архивного дела обладает атрибутами, общими для описаний различных уровней. Для каждого уровня описаны классы: Фонд, Опись, Дело, Документ, унаследованные от Единицы описания.

Ресурс Организация включает организации РАН, научные центры и другие организации. Данные об их сотрудниках сопоставлены ресурсу Персона. Ресурс Проект поддерживает сведения о проектах, выполненных или ведущихся в РАН и других ведомствах. Ресурс Публикация представляет данные о публикациях и научной деятельности.

Метаданные научных библиотек, архивов и музеев

Стандарты - одна из ключевых составляющих инфраструктуры данных институтов памяти. Они задают язык и правила взаимодействия участников, без которых это взаимодействие невозможно. Введение стандартов обеспечивает совместимость на уровне данных и программных средств, позволяет избежать потерь информации и открывает новые возможности по интеграции данных и их совместной обработке.

В библиотечной сфере доминируют системы метаданных, основанные на форматах семейства MARC. Этот формат в значительной степени избыточен, в результате чего возникли проекты систем метаданных METS (Metadata Encoding and Transmission Standard) и MODS (Metadata Object Description Schema), в которых формат MARC адаптируется к современным требованиям информационных систем [3].

В архивной области распространены стандарты метаданных:

- EAD - Encoded Archival Description;
- ISAD(G) - General International Standard Archival Description, Second Edition;
- ISAAR (CPF) - International Standard Archival Authority Record for Corporate bodies, Persons and Families, Second Edition.

Существенное различие во внутренних моделях данных, используемых в различных музеях, библиотеках и архивах, является главной проблемой на пути решения задачи интеграции данных [7]. Перечисленные выше стандарты не поддерживаются или поддерживаются частично [3]. Для преодоления этой проблемы в решаемой задаче интеграции данных было предложено участникам экспортировать метаданные из своего внутреннего формата в формат на базе Dublin Core с использованием синтаксиса XML, так как во внутренних используемых форматах удается выделить общую часть, которая ложится в рамки предложенного формата.

Профиль метаданных LibMeta

Профиль метаданных СУЭБ LibMeta построен на базе библиотечного профиля ЕНИП (Единое научное информационное пространство) РАН [5]. ЕНИП разрабатывается в рамках программы создания и объединения информационных систем подразделений и научных институтов РАН для удовлетворения потребностей научных сотрудников как в части поиска информации, так и в представлении собственной информации в сети Интернет. Одними из наиболее важных составляющих ЕНИП являются схемы метаданных, формализованные с помощью стандартов Semantic Web - RDF [8] /RDFS [9] /OWL [10].

Существенным недостатком многих схем метаданных электронных библиотек является то, что они работают лишь с так называемыми документоподобными объектами, не выделяют другие виды важных объектов, например, персоналии, организации, конференции и т.п. Встретив упоминание персоны в одном месте, невозможно точно установить соответствие с ее упоминанием в другом месте. Даже идентифицировав персону, как пра-

вило, нет возможности получить документы, связанные только с ней. Это обусловлено тем, что метаданные рассматриваются как нечто, связанное только с документом.

В связи с этим в профиле метаданных ЕНИП для электронных библиотек активно используются ресурсы, представленные в основном профиле и некоторых его расширениях, такие как Организации, Персоны и т.д. Тем не менее, центральным остается библиографическое описание публикации, отвечающее за представление метаданных о печатных изданиях.

В целях обеспечения поддержки различных уровней детализации информации о публикациях, необходимых различным приложениям, библиографическая специализация разделена на базовую и расширенную подсхемы, а также выделяется академическая подсхема, отражающая специфику научных публикаций. Уже на базовом уровне требуется структурировать информацию обо всех вышестоящих библиографических уровнях для каждой публикации. Например, для описания ряда статей в журнале необходимо описать сам журнал как издание сводного уровня, далее описать интересующие выпуски этого журнала как издания монографического уровня и, наконец, сами статьи как издания аналитического уровня. И статья, и выпуск, и журнал как таковой являются полноценными структурированными ресурсами, описываемыми лишь единожды и связываемыми с помощью URI-ссылок.

Такой структурированный подход требует некоторого усилия со стороны систем с «планарным» описанием публикаций. Однако структуризация информации обо всех библиографических уровнях необходима и крайне важна для схем ЕНИП. Она позволяет избежать дублирования информации, эффектов наличия опечаток в названиях группирующих выпусков, серий и пр., позволяет представить пользователю информацию в целостном и непротиворечивом виде.

Общая схема профилей метаданных, применяемых в СУЭБ LibMeta, а также основных сущностей в данных профилях приведена на **рис. 1**.

К основным типам данных, представленных в СУЭБ LibMeta, относятся Публикации, Персоны (авторы), Предметы. Сближение задач электронных библиотек, архивов и музеев выдвигает требование стандартизации метаданных физических музейных предметов и их мультимедийных (фото-, видео-, аудио-) представлений. В связи с этим в СУЭБ LibMeta разработаны дополнительные прикладные профили поддержки музейной деятельности и мультимедийных представлений.

В интерфейсе администратора системы появляется возможность определять дополнительные свойства предмета, при этом в интерфейсах ввода и вывода данных создаются представления соответствующих полей. Введенные значения дополнительных полей выдаются в полных сведениях о предмете, но поиск по ним не производится. Таким образом, администратор может добавить такие свойства, как Количество предметов, Автор описания, География, Размеры, Возраст, Способ поступления, Препараты и т.п.

Для обеспечения цифровых представлений публикаций, музейных объектов, а также мультимедийных изображений коллекций, фотографий персон и т.п. разработан дополнительный прикладной профиль Расширенной поддержки хранения данных, в котором вводится ряд новых сущностей. Основные из них - класс Медиа-объект, предназначенный для описания медиа-объекта как единого целого, состоящего из частей данных с различной функциональной нагрузкой, и класс Часть медиа-объекта, позволяющий в пределах одного целого медиа-объекта, например, музейного предмета, иметь несколько частей с различной функциональной нагрузкой, такие как фотографии с разных сторон, видеоролик, сопроводительные информационные документы и т.п. В класс Ресурс, являющийся суперклассом для всех основных объектов онтологии, вводится свойство Медиа-представление. Таким образом, одно или несколько мультимедийных представлений может сопровождать любой объект информационной системы, наследуемый от Ресурс.

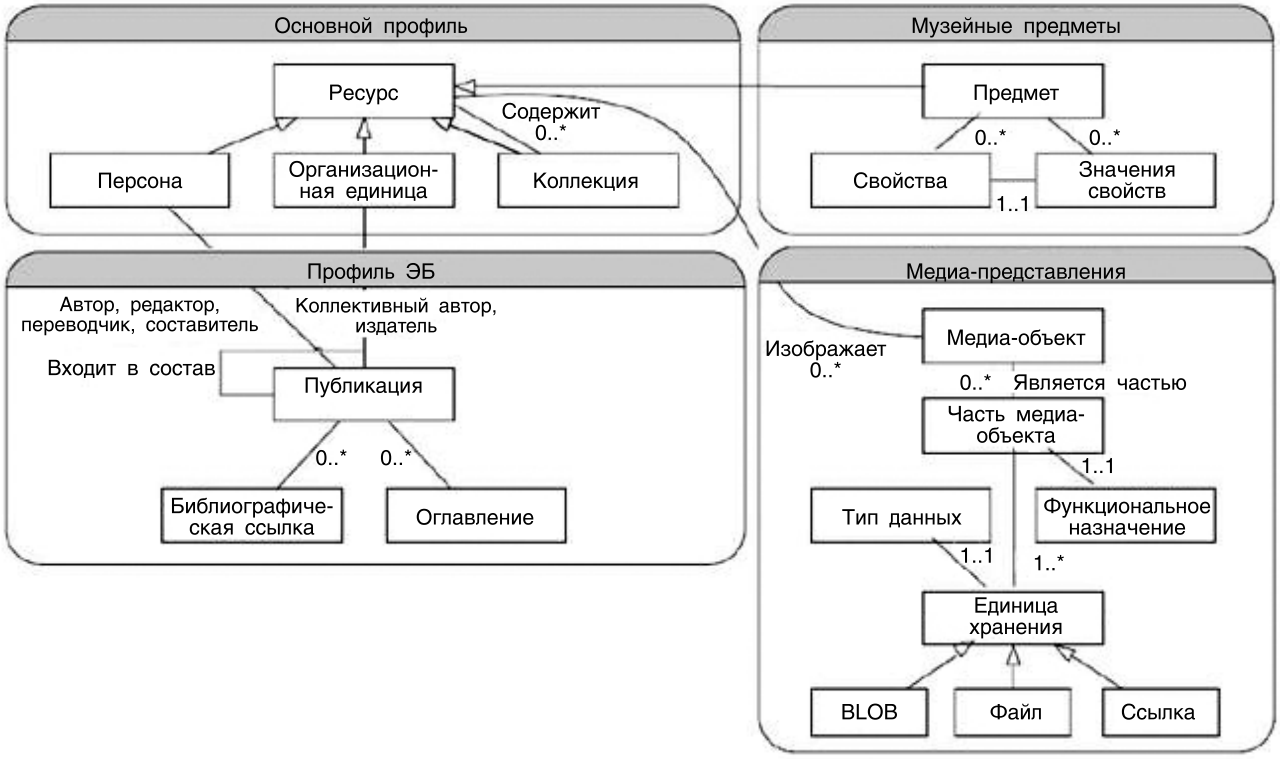

Рис. 1.

В основном профиле метаданных ЕНИП предусмотрена поддержка коллекций, однако требования цифровых библиотек, в особенности с поддержкой хранения музейных предметов, не позволяют их полноценно использовать. В связи с этим базовый профиль дополняется коллекциями со следующими атрибутами: Название, Тип коллекции (элемент словаря), Ключевые слова, Описание, Администратор (ссылка: Персона), Количество элементов в коллекции, Место хранения, Примечание, Элементы коллекции (ссылка: Ресурс). Коллекции такого рода позволяют хранить классические ресурсы (архивные, музейные) и иметь любые вложенные наборы объектов (выставочные, выездные, по хранению и пр.).

Функциональность LibMeta

К основным функциональностям системы «LibMeta» относятся:

- каталогизация, сбор, поиск метаданных объектов электронных фондов институтов памяти;
- каталогизация, сбор, поиск метаданных об ученых, публикациях;
- работа с коллекциями позволяет просматривать, редактировать и выполнять поиск по коллекции;
- работа с наборами дополнительных атрибутов дает возможность создавать наборы атрибутов, назначать их некоторому объекту;
- работа с медиа-объектами позволяет просматривать и редактировать медиа-объекты;
- хранение и просмотр отсканированных текстов;
- ведение словарей и классификаторов, которые могут быть использованы для организации тематического поиска;
- импорт метаданных из внешних систем.

Интерфейс системы представлен веб-порталом, поэтому основным методом доступа пользователя к информации является обычный доступ к веб-страницам портала через любой распространенный браузер. Ядро системы предоставляет следующие возможности: управление статическим содержанием; хранение объектов системы (представленных RDF-тройками) в реляционных СУБД и их пакетную загрузку; индексирование и полнотекстовый поиск; обеспечение безопасности системы, управления новостями.

1.1 Поиск по метаданным

При нажатии пункта меню «Музейные экспонаты» в разделе «Атрибутный поиск» основного меню будет выдана страница поиска метаданных ресурсов. На этой странице можно заполнить нужные поисковые поля. При нажатии кнопки «Поиск» выполняется поиск ресурсов, и результат выдается в формате «Каталог ресурсов». Можно также войти через пункт меню «Тематические разделы» в раздел - будет выдан список тематических разделов с указанием наличия для каждого пункта количества ресурсов с данной тематикой.

При нажатии на некоторый раздел выдается список его ресурсов в формате «Каталог ресурсов» с поисковой формой. Для просмотра основных ресурсов можно выполнить их поиск или нажать пункт меню «Каталог ресурсов» в разделе «Тематические разделы» и перейти на страницу Каталога ресурсов.

1.2 Добавление метаданных

При входе в регистрацию музейных предметов пользователь попадает на страницу ввода основных идентификационных данных о предмете: название, дополнительное название, ключевые слова, abstract, сохранность, количество образцов, размеры, масса, возраст, дата поступления, способ поступления, дата сбора, автор сбора, место сбора, описан в публикации. Идентификаторы и атрибуты можно добавить после сохранения нового объекта.

В разделе «Администрирование» основного меню имеется вход «Загрузка». Здесь можно указать и загрузить файл в стандартном RDF/XML виде.

Поиск, просмотр и регистрация дополнительных ресурсов выполнены стандартными средствами ИС «НИ РАН» и подробно описаны в документации и публикациях по этой системе [4].

Общая архитектура СУЭБ LibMeta

Система управления электронной библиотекой LibMeta включает в себя следующие функциональные подсистемы:

- Подсистема работы с метаданными об ученых, публикациях, музейных объектах позволяет просматривать, редактировать, а также производить поиск информации об ученом, публикации, музейном объекте.
- Подсистема работы с коллекциями позволяет просматривать, редактировать и выполнять поиск по коллекции.
- Подсистема работы с наборами дополнительных атрибутов дает возможность создавать наборы атрибутов, назначать их некоторому музейному предмету.
- Подсистема работы с медиа-объектами позволяет просматривать и редактировать медиа-объекты.
- Подсистема хранения и просмотра отсканированных текстов дает возможность просматривать подряд страницы издания, переходить на любую заданную страницу (в том числе на предыдущую, на последующую, на страницу с заданным номером), просматривать оглавление издания с возможностью перехода на нужный раздел, возможностью просмотра страниц в увеличенном масштабе, выполнять разворот иллюстраций на 90°.
- Подсистема управления структурой статического наполнения портала.
- Подсистема управления группами и пользователями.
- Подсистема управления новостями.
- Подсистема ведения словарей и классификаторов, которые могут быть использованы для организации тематического поиска.
- Подсистема пакетной загрузки данных позволяет загружать данные в формате RDF/XML в соответствии с онтологической моделью метаданных LibMeta.
- Подсистема полнотекстового поиска информации об ученых, публикациях, музейных объектах, коллекциях и медиа-объектах.
- Подсистема импорта метаданных, а также подготовленных электронных изданий и их оглавлений из внешних систем.

Интеграция СУЭБ LibMeta с другими информационными системами

Основой для описания схем метаданных в ЕНИП и LibMeta служат технологии Semantic Web. В Semantic Web широко используется язык RDF, а также его специализация для описания онтологий - OWL. Логичным является выбор RDF как языка обмена метаданными между системами. Кроме того, для интеграции с универсальными агрегаторами в СУЭБ LibMeta поддерживается взаимодействие по протоколу OAI-PMH [5], базирующееся на использовании стандарта Dublin Core [11].

В системе создан универсальный модуль загрузки метаданных в произвольном XML-формате в соответствии с протоколом OAI-PMH. С определенной периодичностью интеграционный модуль запрашивает вновь созданные или измененные метаданные из удаленного хранилища по протоколу OAI-PMH. В первую очередь проверяется URI (Unified Resource Identifier) получаемых метаданных. Если метаданные с указанным URI уже представлены в системе, то выполняется XSLT-преобразование [12] (метаданные приводятся к внутреннему RDF/XML/-формату СУЭБ LibMeta) и производится загрузка в режиме «дозапись».

При загрузке в режиме «Дозапись новых данных поверх существующих» для каждого свойства, загружаемого из RDF/XML, все прежние значения этого свойства

стираются и заменяются на значения из RDF/XML. При этом значения тех свойств, которые были указаны в базе, но отсутствуют в RDF/XML, оставляются неизменными. Такой режим загрузки обеспечивает корректную инкрементную «дозапись» данных поверх существующих. Если метаданных с указанным URI в системе нет, то они являются новыми и также должны быть загружены. Однако в силу того, что СУЭБ LibMeta представляет собой единый интеграционный узел, метаданные, соответствующие некоторому информационному ресурсу, могут быть получены ранее из другого источника. Для того чтобы в СУЭБ LibMeta не возникало дубликатов, используется вспомогательный модуль автоматической проверки на дубликаты [13]. Этот модуль предоставляет возможность указания параметров проверки. Так, например, может быть указано, что два ресурса представляют одну и ту же публикацию, если у них совпадают названия и списки авторов. Если есть предположение о том, что загружаемые метаданные уже хранятся в системе, источнику метаданных отправляется информация о схожих метаданных (и их уникальных идентификаторах), находящихся в СУЭБ LibMeta.

Заключение

Портал «Libmeta» является результатом научно-исследовательской работы, проводимой при поддержке РФФИ, и объединяет АИС нескольких участников, а именно НПБ им К.Д. Ушинского, Государственного Дарвиновского музея, ГПНТБ России, Архива РАН, БЕН РАН. Имеется возможность подключения новых участников. Использованы ключевые, международные стандарты, применяемые при разработке СУЭБ, ее общая архитектура и основные элементы. В проекте реализованы средства интеграции приложений с разными источниками/каталогами метаданных/данных, сервис директорий метаданных, унифицированный интерфейс поиска данных. Особое внимание уделено стандартам и методам реализации платформы для создания распределенной среды интеграции неоднородных источников данных электронных фондов институтов памяти, а также практике использования СУЭБ в некоторых работах, выполняемых совместно с организациями РАН.

Для ознакомления портал доступен по адресу **http://libmeta.ru.**

Литература:

1. Электронный ресурс http://www.europeana.eu/portal/

2. Электронный ресурс http://www.socionet.ru/

3. Антопольский А.Б. Вопросы интеграции библиотек, архивов и музеев по научному наследию // Информационное обеспечение науки: новые технологии, 2011.

4. Бездушный А.А. Информационная Web-система «Научный институт» на платформе ЕНИП / Бездушный А.А., Бездушный А.Н., Нестеренко А.К., Серебряков В.А, Сысоев Т.М., Теймуразов К.Б., Филиппов В.И. - М.: Вычислительный центр РАН, 2007.

5. Бездушный А.А. Интеграция метаданных Единого Научного Информационного Пространства РАН / Бездушный А.А., Бездушный А.Н., Серебряков В.А, Филиппов В.И. - М.: ВЦ РАН, 2006. - 238 с.

6. Захаров А.А., Серебряков В.А. Система управления электронными библиотеками LibMeta // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды XII Всероссийской научной конференции RCDL'2010. - Казань: КФУ, 2010. - 28 с.

7. А.Б. Антопольский, А.А. Каленкова, Н.Е. Каленов, В.А. Серебряков, А.Н. Сотников. Принципы разработки интегрированной системы для научных библиотек, архивов и музеев // Информационные ресурсы России. - 2012. - № 1. - С. 2-6.

8. Resource Description Framework (RDF) Model and Syntax, W3C Recommendation, 2004. - http://www.w3.org/TR/rdf-primer/.

9. Resource Description Frame-work (RDF) Schema Specification. W3C Candidate Recommendation, 2004. - http://www.w3.org/TR/rdf-schema.

10. OWL Web Ontology Language Semantics and Abstract Syntax, 2004. - http://www.w3.org/TR/2004/REC-owl-semantics-20040210/.

11. The Dublin Core Metadata Element Set: an American national standard // Dublin core metadata initiative. - http://dublincore.org/documents/dces/.

12. Open archives initiative protocol for metadata harvesting // http://www.openarchives.org/pmh.

13. XSL Transformations (XSLT) Version 2.0, W3C Recommendation // W3C http://www.w3.org/TR/xslt20/.

14. Атаева О.М., Шиолашвили Л.Н. Методы очистки интегрируемых данных // Современные проблемы фундаментальных и прикладных наук: Труды XLIX научной конференции. - М.: МФТИ. - 2006.

НАША ИНФОРМАЦИЯ

Конференция «Система обеспечения российских организаций научно-технической информацией в электронном виде. Итоги и перспективы проекта МОН»

С 6 по 8 ноября 2012 года в Санкт-Петербурге состоится конференция «Система обеспечения российских организаций научно-технической информацией в электронном виде. Итоги и перспективы проекта МОН».

Место проведения:

Санкт-Петербург, Кронверкский пр., 49.

Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики (НИУ ИТМО);

Санкт-Петербург, г. Пушкин, ул. Радищева, 4.

Конгресс-комплекс «Особняк В.П. Кочубея».

Организаторы:

Министерство образования и науки Российской Федерации;

Некоммерческое партнерство «Национальный Электронно-Информационный Консорциум» (НП «НЭИКОН»);

Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики (НИУ ИТМО).

Информационные спонсоры:

www.neicon.ru

Основные задачи конференции:

Отчет о ходе выполнения и обсуждение перспектив развития проектов Министерства образования и науки РФ в области обеспечения научной информацией в электронной форме.

Обсуждение государственных инициатив в направлении интенсификации российских научных исследований.

Тематика конференции:

Реализация Некоммерческим партнерством «Национальный Электронно-Информационный Консорциум» проекта Министерства образования и науки РФ.

Опыт университетов и библиотек в области организации работы с научной информацией в электронной форме.

Формирование новых принципов подписки на электронные источники информации.

Международный и российский опыт организации доступа к архивам наиболее авторитетных информационных источников в электронной форме.

Возможные шаги в направлении интенсификации проведения научно-исследовательских работ в России и их оценки.

Сайт: *http://conf.neicon.ru/index.php/science/mon2012*