Интеграционные возможности Системы управления электронными библиотеками LibMeta

А.А. Каленкова, В.А. Серебряков

Аннотация

В статье представлена Система управления электронными библиотеками (СУЭБ) LibMeta, предназначенная для интеграции библиотечной и музейной информации. Приводится схема метаданных СУЭБ LibMeta, построенная на базе схемы ЕНИП, она включает основной и библиотечный профили, а также профили для работы с музейными и медиа-объектами. Интеграция метаданных из различных источников может привести к дублированию информации: дается подробный алгоритм интеграции метаданных, который позволяет избежать дублирования мета-описаний.

Ключевые слова: цифровые библиотеки, метаданные, интеграция данных.

1. Введение

В последние годы объемы информации в сети Интернет в связи с ее бурным развитием лавинообразно увеличиваются [3]. Несмотря на все большее проникновение технологий Semantic Web [1, 2], ощущается серьезная нехватка средств поиска и каталогизации информации, которые позволяли бы искать ее именно по семантике и связям, а не только по ключевым словам и полным текстам, как это делают универсальные поисковые системы. Одним из способов решения данной проблемы видится появление и все большее распространение различного рода ЭБ (электронных библиотек) [12, 14]. Программы развития ЭБ достаточно активно начали разрабатываться в мире, начиная с 90-х годов прошлого века. К настоящему моменту в мире осуществляется большое количество международных и национальных проектов по созданию электронных библиотек. Активно ведутся работы по выработке, принятию и поддержке международных стандартов в области формирования электронных информационных фондов и процедур доступа к ним.

Особое значение имеет интеграция различных (в том числе и библиотечных) информационных ресурсов, она может быть выполнена таким образом, что в центральной информационной системе будет храниться лишь метаинформация необходимая для навигации и семантического поиска, в то время как сами данные будут располагаться в других информационных системах. Ресурсы, даже хранящиеся в разных системах, представляются связанными друг с другом единой системой навигации. При интеграции метаданных возможно возникновение дубликатов, то есть разные информационные системы могут предоставлять информацию об одном и том же ресурсе. Эти ситуации должны отслеживаться со стороны центральной информационной системы при загрузке, кроме того, люди загружающие метаданные должны получить возможность принимать окончательное решение о том, совпадают ли загружаемые ресурсы с уже существующими в системе ресурсами. Центральной информационной системой, реализующей указанный подход к семантической интеграции, является система управления электронными библиотеками LibMeta [13]. С 2007 года в ВЦ РАН ведутся работы по созданию этой системы.

2. Схема метаданных СУЭБ LibMeta

Профиль метаданных СУЭБ LibMeta построен на базе библиотечного профиля ЕНИП (Единое научное информационное пространство) РАН [11]. 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 [6] / RDFS [7] / OWL [5].

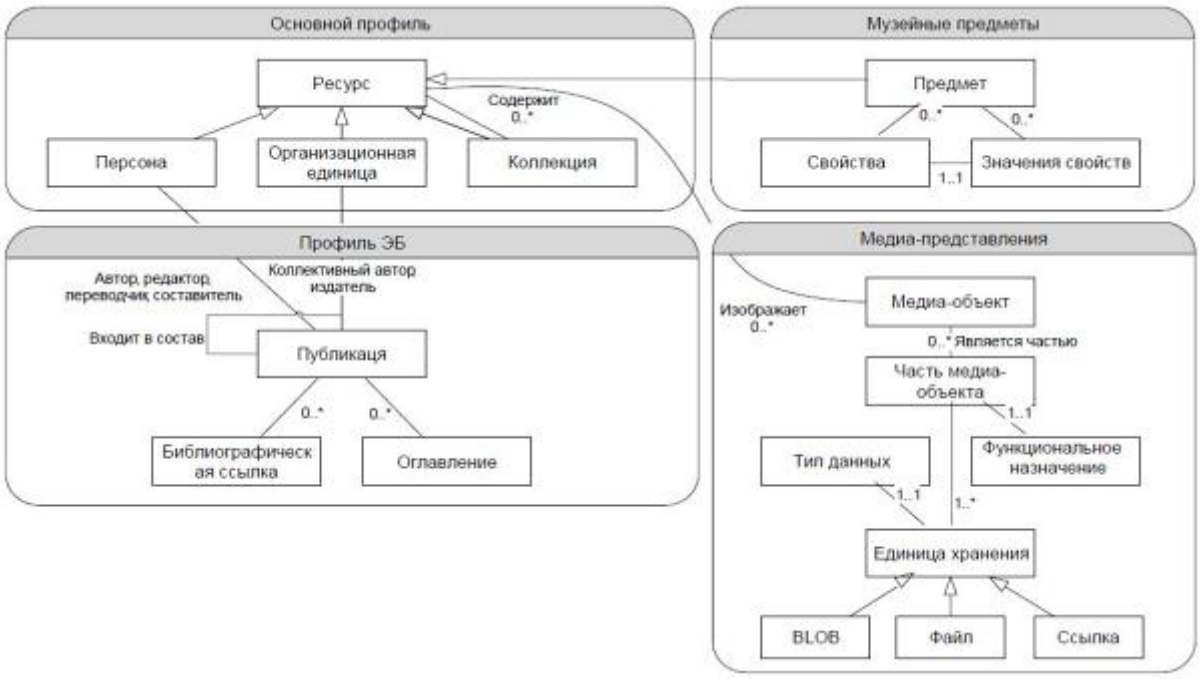

Рис. 1. Схема профилей метаданных СУЭБ LibMeta

Существенным недостатком многих схем метаданных электронных библиотек является то, что они работают лишь с так называемыми документо-подобными 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 правило, 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 (Unified Resource Identifier)-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 рисунке 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 отличие от публикаций, описания музейных объектов могут значительно отличаться в различных музеях и здесь невозможно обеспечить всеобъемлющий набор необходимых свойств. В связи с этим для данных объектов реализуется возможность определения дополнительных свойств в виде связей с двумя вспомогательными объектами: Дополнительные свойства и Значения дополнительных свойств. Соответственно, 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.

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 эти 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.).

3. Общая архитектура СУЭБ LibMeta

Система управления электронными библиотеками 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 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 оглавления издания с возможностью перехода на нужный раздел, возможность 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.

4. Интеграция СУЭБ LibMeta с другими информационными системами

Основой для описания схем метаданных в ЕНИП и LibMeta служат технологии Semantic Web.

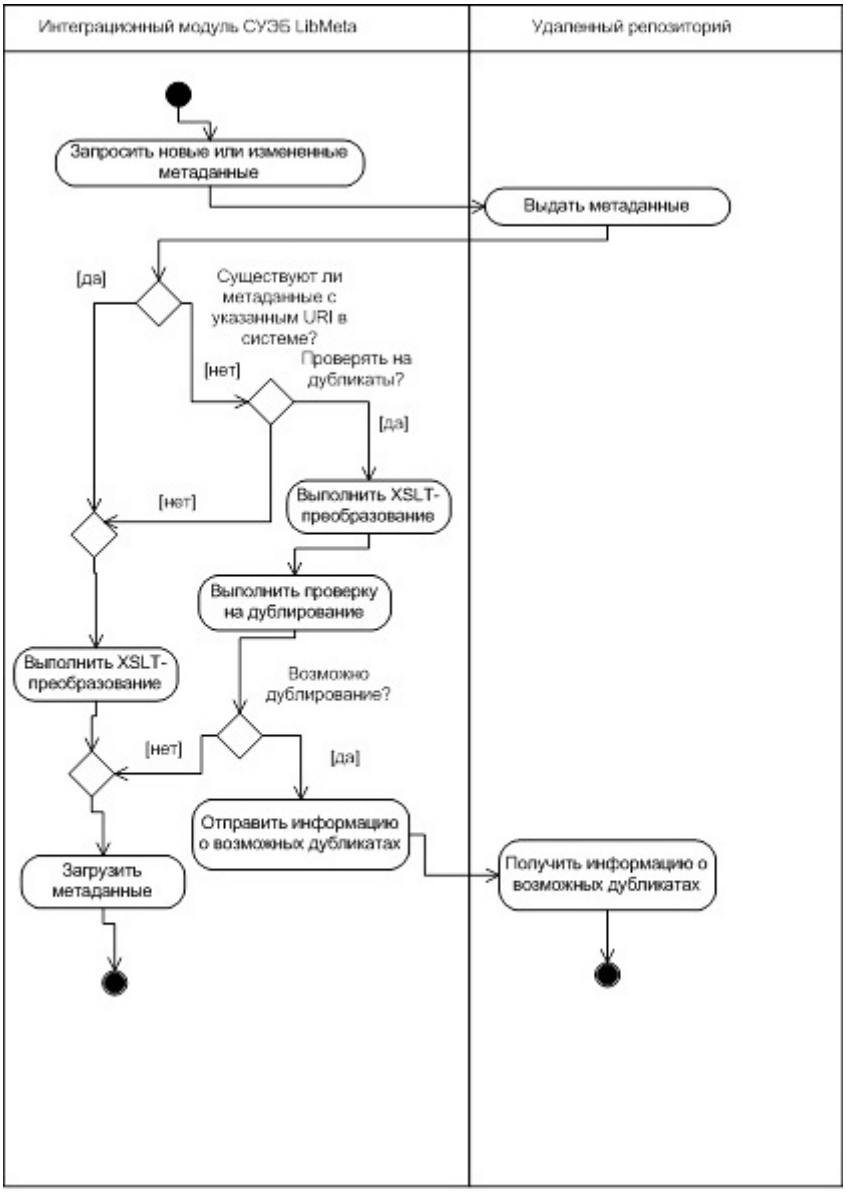

Рис. 2. Алгоритм работы интеграционного модуля СУЭБ LibMeta

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 [4], базирующиеся на использовании стандарта Dublin Core [8]. В системе создан универсальный модуль загрузки метаданных в произвольном XML-формате в соответствии с протоколом OAI-PMH. Алгоритм получения метаданных некоторого ресурса, реализованный в этом модуле, представлен на рисунке 2.

С определенной периодичностью интеграционный модуль запрашивает вновь созданные или измененные метаданные из удаленного хранилища по протоколу OAI-PMH. В первую очередь проверяется URI получаемых метаданных. Если метаданные с указанным URI уже представлены в системе, то выполняется XSLT [9] – преобразование (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 [10]. 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. Источнику сообщается: какие метаданные были загружены, какие метаданные не были загружены в силу ошибок (в этом случае приводится описание ошибок загрузки), какие метаданные имеют сходство с некоторыми метаданными уже находящимися в СУЭБ LibMeta, они не загружаются, и приводится список дубликатов и их URI. При возникновении подозрений на наличие дубликатов на стороне источника определяется, соответствуют ли метаданные одному и тому же информационному ресурсу. Если принимается решение о том, что эти метаданные уже есть в системе, для них устанавливается URI уже загруженных метаданных (тогда при следующей загрузке метаданные в репозитории могут быть дополнены новыми значениями полей),

во всех своих метаданных, ссылающихся на этот ресурс, также устанавливается в качестве значений ссылок выбранный идентификатор. Иначе, для загружаемых метаданных выставляется признак, что они должны быть загружены, несмотря на наличие схожих метаданных в СУЭБ LibMeta, и они попадают в систему при следующей загрузке без проверки на дублирование. Таким образом, интеграционный модуль СУЭБ LibMeta реализует некоторый общий подход к загрузке метаданных из удаленных репозиториев.

Литература

[1] Berners-Lee T. The Semantic Web Revisited / Berners-Lee T., Shadbolt N., Hall W. // IEEE Intelligent Systems. 2006. N. 6.

[2] Berners-Lee T. The Semantic Web / Berners-Lee T., Hendler J., Lassila O. // Scientific Am. 2001. N. 5. P. 34–43.

[3] Gantz J. The Diverse and Exploding Digital Universe: An Updated Forecast of Worldwide Information Growth through 2011/ Gantz J., Chute C., Manfrediz A., et al. // IDC White paper. 2008.

[4] Open archives initiative protocol for metadata harvesting // OAI [Электронный ресурс]. — Режим доступа: http://www.openarchives.org/pmh.

[5] OWL Web Ontology Language Semantics and Abstract Syntax // W3C [Электронный ресурс]. — Режим доступа: http://www.w3.org/TR/2004/REC-owl-semantics-20040210/.

[6] Resource Description Framework (RDF) Model and Syntax // W3C [Электронный ресурс]. — Режим доступа: http://www.w3.org/TR/rdf-primer/.

[7] Resource Description Framework (RDF) Schema Specification. // W3C [Электронный ресурс]. — Режим доступа: http://www.w3.org/TR/rdf-schema.

[8] The Dublin Core Metadata Element Set: an American national standard // Dublin core metadata initiative [Электронный ресурс]. — Режим доступа: http://dublincore.org/documents/dces/.

[9] XSL Transformations (XSLT) Version 2.0, W3C Recommendation // W3C [Электронный ресурс]. — Режим доступа: http://www.w3.org/TR/xslt20/.

[10] Атаева О.М. Методы очистки интегрируемых данных / Атаева О.М., Шиолашвили Л.Н. // Современные проблемы фундаментальных и прикладных наук. 2006.

[11] Бездушный А. Н. Интеграция метаданных Единого Научного Информационного Пространства РАН / Бездушный А. Н., Бездушный А. А., Серебряков В. А., Филиппов В. И. М.:ВЦ РАН, 2006.

[12] Галева И. С. Интернет как инструмент библиографического поиска // Профессия. 2007.

[13] Захаров А.А.Система управления электронными библиотеками LibMeta / Захаров А.А., Серебряков В.А. // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. 2010. – С. 28.

[14] Зацман И. М. Концептуальный поиск и качество информации / М.: Наука, 2003.

Работа выполняется в рамках проекта РФФИ №11-07-00286-а.

Об авторах

Каленкова Анна Александровна – Вычислительный центр им. А.А.Дородницына РАН, г. Москва e-mail: akalenkova@ultimeta.ru

Серебряков Владимир Алексеевич – Вычислительный центр им. А.А. Дородницына РАН, Москва e-mail: serebr@ccas.ru