Персональная цифровая библиотека Libmeta как среда интеграции связанных открытых данных

© О. М. Атаева © В. А. Серебряков
Вычислительный центр им. А.А. Дородницына РАН,
Москва
oli@ultimeta.ru serebr@ultimeta.ru

Аннотация

В статье описывается семантическая электронная библиотека Libmeta, ресурсы которой могут быть обогащены за счет использования данных из источников, расположенных в LOD. Связывание происходит посредством онтологии предметной области, которая задается пользователем и определяет его область интереса. Затрагиваются проблемы интеграции ресурсов библиотеки в LOD и создания поисковых запросов по источникам данных, а также обсуждается использование спецификаций и технологий из стека LOD в рамках одной системы.

* Работа выполнена при поддержке РФФИ – проект № 14-07-00058 А.

1 Введение

Последнее десятилетие наблюдается бурное развитие технологий Semantic Web и активное развитие сообщества, поддерживающего Linked Open Data (LOD). Основная идея LOD заключается в решении задач интеграции данных, представленных в сети, для чего предлагается представить информацию в формализованном виде, что делает ее доступной для машинной обработки.

Развитие технологий Semantic Web и популярность идеи LOD оказали влияние и на электронные библиотеки, которые трансформируются и превращаются в центры данных, вокруг которых формируется сообщество заинтересованных экспертов и пользователей, принимающих активное участие в их развитии. При консорциуме W3C была создана рабочая группа под названием Linked Library Data, которая выработала рекомендации по связыванию библиографических данных с использованием стандартных семантических технологий RDF, SPARQL, OWL. Появление семантических технологий вызывает необходимость разработки новых подходов к созданию электронных библиотек и расширяет возможности их использования.

Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.

2 Эволюция библиотек

Развитие информационных технологий в XX веке и их использование в библиотеках привело к появлению нового типа библиотек [16].

2.1 Электронные библиотеки

Электронные библиотеки возникли достаточно давно и представляют собой набор документоподобных ресурсов и их библиографии, в доступных для компьютеров форматах, а также сопутствующих услуг для их хранения и поиска. При этом в таких библиотеках не выделялись другие виды важных объектов, например, персоналии, организации 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 рассматривались как нечто, связанное только с документом.

2.2 Цифровые библиотеки

Цифровые библиотеки представляют собой информационные системы, которые обеспечивают задачи коллекционирования, хранения и навигации по разнообразным электронным документам, как хранящимся в самой системе, так и доступных по сети. Термин «цифровые библиотеки» часто рассматривается как синоним термина «электронные библиотеки», тогда как цифровые библиотеки являются продуктом следующего этапа развития электронных технологий и исследований в области электронных библиотек, использование результатов которых позволило расширить функциональность электронных библиотек, превратив их в «цифровые».

2.3 Семантические цифровые библиотеки

Использование семантических технологий значительно расширяет функциональность библиотек: данные лучше структурированы,

выделены связи между ними, улучшается поиск, появляется возможность интегрировать данные различных типов: персоны, ресурсы, пользователи. Обеспечивается интероперабельность с другими системами, не обязательно являющимися библиотеками, так как основной задачей семантических технологий остается предоставление метаданных в машиночитаемом формате. Онтологии играют основную роль для решения задач, вызванных структурными различиями существующих систем и семантическими различиями стандартов метаданных.

2.4 Персональные семантические цифровые библиотеки

Мы выделяем персональные семантические цифровые библиотеки, наполнение которых индивидуально для каждого пользователя системы и выполняется в полуавтоматическом режиме из разнородных источников данных, интегрированных в облако LOD. Будем далее для краткости называть их персональными открытыми цифровыми библиотеками или ПОЦБ. Типы информационных ресурсов и их структура определяются пользователем, исходя из своих интересов, то есть пользователь описывает интересующую его предметную область, определяя тематическое наполнение библиотеки.

Данная статья является продолжением нашей предыдущей работы [1], в которой представлена общая схема системы, выделены ее основные модули и дана характеристика каждого из них. Основное развитие системы произошло в направлении поиска источников связанных данных с использованием технологий из стека LOD. В следующих разделах привидено описание этого модуля и детализация его функций, а также кратко описаны первые практические результаты.

Основная задача системы заключается в предоставлении пользователю унифицированного представления для возможности автоматизированного извлечения интересующей его информации по определенной предметной области.

Представление ресурсов библиотеки в виде связанных данных расширяет функциональность семантических цифровых библиотек, давая возможность:

- включения дополнительных элементов описания данных информационных ресурсов,
- полного или частичного обновления данных из источников,
- использования интерфейсов для создания запросов к интегрированным в LOD источникам данных на основе SPARQL,
- включения в описания ресурсов других типов информации.

Одна из задач, которая решается в ПОЦБ, – это реализация интеграции набора данных в пространство LOD с использованием онтологии предметной области информационных ресурсов, т.е. автоматизированное обнаружение новых наборов данных и, по возможности, установка и поддержка связей с элементами данных из этих наборов данных с уже имеющимися ресурсами в репозитории библиотеки, обеспечивая одновременно рекомендуемую проектом LOD функциональность в рамках одной системы.

3 Источники данных

Мы подразделяем источники данных на два типа: внешние и внутренние. Внешними мы называем те источники, которые интегрированы в LOD, и данные которых представлены в RDF и доступны нам с использованием SPARQL. Для своих практических целей мы использовали такие известные источники в LOD, как DBpedia [3], Europeana [4]. Внутренние источники могут представлять собой любой другой тип источника данных, который не интегрирован в LOD. На практике в качестве внутренних источников мы использовали другие библиотеки, которые предоставляли доступ к своим данным по протоколу OAI-PMH.

3.1 Внешние источники

Данные из источников LOD хорошо структурированы и обычно доступны через SPARQL точку доступа для поисковых запросов. Так как одним из принципов LOD является использование URI, по которым можно получить по HTTP информацию в стандартном формате, то для доступа к информации определенного ресурса пользователь может использовать только этот URI.

Основной задачей подсистемы подключения внешних источников является создание и поддержка отображения онтологии предметной области на схему источника данных, посредством которого пользователь получит возможность автоматического мониторинга для последующего связывания имеющихся данных в системе с новыми данными по определенным запросам в терминах своей онтологии. При этом в системе при импорте может сохраняться лишь внешний URI ресурса.

3.2 Внутренние источники

Несмотря на активное развитие LOD, нельзя игнорировать источники данных, которые в него еще не интегрированы и при этом содержат огромный объем полезных данных. По этой причине в нашей системе реализован блок поддержки протокола OAI-PMH, который широко используется в библиотечной среде для обмена метаданными. Основным его недостатком, с точки зрения извлечения информации опираясь на принципы LOD, является то, что для доступа к информации о ресурсе нужно обладать специальными знаниями о протоколе, при этом знание идентификатора OAI, который используется в таком источнике для представления информации о ресурсе, не сильно облегчает поиск этих данных. Например,

OAI идентификатор ресурса *42041024*, на портале «Научное наследие Росии», для пользователя не обладающего специальными знаниями не позволит найти полезной информации, тогда как идентификатор из источника LOD http://dbpedia.org/page/Mikhail_Lomonosov интуитивно понятен и позволяет получить доступ к полезной информации о ресурсе, а также к связанным с ним ресурсам. Таким образом поддерживая этот протокол, мы внутри нашей системы решаем задачу формального предоставления и интеграции этих данных в соответствии с принципами LOD, при этом сохраняя информацию о первоначальном источнике, одновременно позволяя решать задачи связывания данных с другими источниками из облака LOD в рамках системы.

В работе [2] предлагается улучшенная версия этого протокола, которая является развитием протокола в сторону поддержки связанных данных.

4 Функциональность ПОЦБ

К основной функциональности системы, реализующей ПОЦБ относятся:

- функции атрибутного поиска;
- функция выделения неявных связей между ресурсами по их описаниям;
- функция работы с коллекциями;
- создание/просмотр/редактирование/объединение/вложенные коллекции;
- функция отображения онтологии ИД;
- функция детализации, которая обеспечивает преобразование в подзапросы, соответствующих различным ИД;
- функция для выполнения запросов и обработки результатов и предоставления окончательного результата пользователю;
- функция автоматического мониторинга ИД на наличие новых/измененных данных;
- создание словарей, классификаторов, тезаурусов;
- редактирование элементов;
- поддержка («гибкой») классификации ресурсов;
- поддержка настройки уровней доступа к различным ветвям тезауруса.

Исходя из определения источников данных ПОЦБ и перечня функций системы, можно выделить «внутренние» функции, т.е. те, которые оперируют данными в рамках системы и интегрируют данные из «внутренних» источников и фактически определяют обычную семантическую библиотеку. «Внешние» функции обеспечивают подключение и извлечение данных из LOD и позволяют задать тематическое наполнение библиотеки и установить связи, таким образом задавая фактически определение ПОЦБ.

5 Онтология ПОЦБ

Онтология ПОЦБ разработана в общем виде без привязки к конкретным методам и способам реализации семантических цифровых библиотек [1], [5].

Фактически общая онтология ПОЦБ состоит из двух онтологий:

1) онтология СЭБ, построенная на основе онтологии информационных систем, включающая в себя основные понятия, необходимые для обеспечения основной функциональности библиотеки, такие как ресурс, пользователь, коллекция, словарь, классификатор, запрос, источник и т.д.

2) онтология и тезаурус предметной области, для которой пользователь определяет ее понятия, их тип, структуру, совокупность словарей и классификаторов, которые представляют тезаурус предметной области, который обеспечивает доступ неквалифицированных пользователей, решающих задачи поиска информации, к знаниям предметной области в разных источниках. Эта онтология позволяет:

- выработать и зафиксировать общее понимание области знания;
- представить знания в удобном для обработки автоматизированными подсистемами виде, обеспечить возможность получения и накопления новых знаний, а также представить возможность многократного использования знаний

Тезаурус же обеспечивает терминологическую поддержку и помогает пользователям сформулировать запрос к системе, в том числе, подобрать правильные ключевые слова для описания искомого результата, имеющихся данных и контекстной информации.

Тезаурус необходим для навигации и для автоматического уточнения и расширения запроса, введенного пользователем, посредством использования зафиксированных в тезаурусе связей между терминами. Например, в частном случае, в качестве предметной области рассматривается онтология из работы [6] со всем набором словарей и классификаторов. Данные, представленные этой онтологией, представляют собой численные значения теплофизических свойств для различных веществ в разных условиях и их библиографии.

Основным классом, поддерживаемым в онтологии СЭБ, является класс *информационный ресурс*, подклассами которого являются такие классы ресурсов как *публикация*, *персона* и т.д. Подключаемые классы предметной онтологии могут являться как подклассами класса *информационный ресурс*, так и расширять структуру подклассов этого класса. Таким образом онтология предметной области одновременно может расширять список информационных ресурсов системы, а также дополнять и расширять структуру информационных

ресурсов. Для поддержки такой интеграции онтологии реализован отдельный модуль поддержки различных типов связей определен минимальный словарь этих связей. Такой подход к созданию онтологии системы позволяет конкретизировать область интересов в рамках конкретной персональной библиотеки.

6 Поиск по источникам данных

Поисковые системы, ориентированные на источники, интегрированные в LOD, такие как Sig.ma, Falcons, и SWSE, обеспечивают поиск на основе ключевых слов, ориентированный на использование той же парадигмы, что и существующие лидеры рынка, такие как Google и Yahoo. Пользователю предоставляется окно поиска, в котором он может ввести ключевые слова, связанные с предметом или темой, в которых он заинтересован, и приложение возвращает список результатов, которые могут (или нет) иметь отношение к запросу. Фактически это поиск по вхождению слова в любой элемент описания. Поиск же данных в источниках предполагает, что пользователь знает структуру данных

В работе [8] представлена система поиска LOQUS в репозиториях LOD на основе высокоуровневой онтологии, на которую отображается схема подключаемого источника данных (ИД). Эта онтология составлена на основе высокоуровневой онтологии, которая содержит наиболее общие и самые абстрактные концепты, имеет исчерпывающую иерархию фундаментальных понятий (около 1 тыс.), а также набор аксиом (примерно 4 тыс.), определяющих эти понятия. Каждому концепту определен идентификатор или обобщающее понятие из LOD. Онтология так же, как и в нашем подходе, используется для трансляции SPARQL запросов пользователей в интегрированные ИД. Но недостаточный уровень концептуализации понятий не позволяет в достаточной мере сконцентрироваться на определенной предметной области.

С другой стороны задача автоматизированного поиска релевантных источников данных осложняется тем, что чаще всего информация о связях между ними проставляется в основном на уровне данных с помощью связей sameAs, seeAlso. Даже простой анализ связей sameAs, seeAlso на уровне найденных данных позволит выявить эквивалентные классы, ранее не определенные связи между разными источниками или новые источники. Описание связей на уровне схем затем можно использовать при формировании запросов к источникам данных.

До недавнего времени связи между источниками на уровне схем описывались гораздо реже. В последние несколько лет эта задача решается с введением и активным распространением спецификации VOID [7] для описания источников RDF данных, в которой предоставляется информация о связанных источниках данных. VOID описание содержит информацию об используемых словарях, статистическую информацию о том, сколько ресурсов того или иного типа или значений определенных свойств используются во множестве. При создании словаря VoID была сведена к минимуму необходимость создания новых свойств и классов, путем использования существующих словарей. Например, для описания статистической информации используется словарь SCOVO. На основе этой информации можно делать вывод о релевантности источника тому или иному запросу или предметной области.

В рассматриваемой системе VoID описание набора данных в хранилище генерируется с помощью D2R Server [15]. В сгенерированное описание не попадает информация о подключенных источниках данных и статистика по имеющимся с ними связям. Для включения этой информации были использованы правила, по которым осуществляется поиск связанных данных [12]. Полученное описание в рамках используемой системы позволяет формировать распределенные запросы к подключенным источникам данных в терминах онтологии, используемой в этой системе. Используя VoID описание, запросы из системы транслируются в термины уже источников данных. Также это описание применяется для отображения обобщенного результата поиска.

7 Общая схема подключения источников данных

На рисунке 1 представлена общая схема подключения различных источников данных с использованием технологий из стека проекта LOD

Доступ к данным Libmeta осуществляется через ее общую онтологию, которая, как было сказано, состоит из: а) онтологии семантической библиотеки, б) онтологии предметной области, которая задает тематическое направление информационных ресурсов. При этом D2R Server [15] использует онтологию Libmeta для создания SPARQL точки доступа к ее данным. Используются правила, которые задаются для каждого подключаемого источника (правил может быть несколько), с помощью которых осуществляется поиск и сохранение связей между данными Libmeta и источником из LOD. Для задания правил связывания используется фреймворк SILK. Правила описываются в соответсвии с требованиями SILK и хранятся в определенном для каждого источника месте. После описания правила и указания его расположения все действия по запуску и анализу результатов работы SILK выполяются программно, ля этого используется соответсвующая задаче версия фреймворка.

При каждом подключении нового источника или обновлении набора связей уже подключенных нужно обновлять VoID описание множества данных Libmeta, анализируя полученный набор ссылок и

правила, по которым они выполнялись. Это позволит обновить статистическую и структурную части VoID, необходимых для использования при формировании запросов в терминах общей онтологии и их преобразования в запросы к релевантным источникам в соответствующим им терминах.

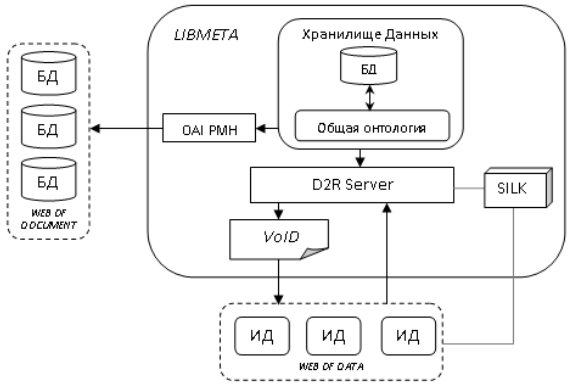

Рисунок 1

Libmeta также исторически поддерживает обмен данными по протоколу OAI-PMH с библиотеками, неинтегрированными в LOD, выступая агрегатором, который интегрирует их данные в LOD.

8 Текущее состояние работ

В рамках создания первой версии ПОЦБ был реализован проект по созданию стандартизированной и децентрализованной среды управления информацией электронных фондов Libmeta [10]. В проекте реализованы средства интеграции приложений с разными источниками/каталогами метаданных/данных, сервис директорий метаданных, унифицированный интерфейс поиска данных.

Существенное различие во внутренних моделях данных, используемых в различных музеях, библиотеках и архивах, является главной проблемой на пути решения задачи интеграции данных [9]. 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. В системе используется универсальный модуль загрузки метаданных в произвольном XML-формате в соответствии с протоколом OAI-PMH.

Основная коллекция метаданных была получена из библиотеки (тип источника внутренний) «Научное Наследие России» [10]. Для интеграции данных в LOD в качестве внешних источников было проведено связывание с данными DBpedia по авторам, а для связывания музейных экспонатов был проведен эксперимент с данными из Europeana.

Для каждого ресурса Libmeta может быть получено его представление, удовлетворяющее модели Europeana Semantic Elements (ESE) [14], которое определяет ряд обязательных элементов метаданных.

Для мониторинга новых данных и установления связей с внешними источниками данных в рамках системы используется SILK Framework [12]. Для установления связей необходимо указать источник данных, правила доступа к данным и правила связывания. Вся эта информация была написана в виде конфигурационного файла на языке SILK LSL.

Сейчас проводятся работы по связыванию данных с авторитетными файлами VIAF [13]. Это проект, который объединяет все значимые библиотеки, интегрирующие свои данные в LOD.

9 Заключение и дальнейшие работы

Разрабатываемая ПОЦБ предполагает поддержку функциональности, рекомендуемую проектом LOD, а именно: средства для представления информации из различных источников как для установления, так и для поддержки связей между RDF-ресурсами, как внутренними, так и внешними, т.е. предполагает осуществление полного цикла интеграции набора данных в пространство LOD.

Основные преимущества реализации принципов LOD в Libmeta:

- Связность. Подключение источников, не обязательно библиотек;
- Машиночитаемость.Представление в RDF, использование общепринятых словарей и онтологий;
- Доступность. Доступные для свободного использования всеми пользователями без каких-либо ограничений в виде авторских прав.

Использование онтологии предметной области позволит не только включать другие типы ресурсов в библиотеку, но и уточнять и включать в библиотеку описания внутренней структуры информационных ресурсов нужной детализации, обращаясь за данными к источникам, которые раньше с трудом могли использоваться в рамках интеграции ресурсов электронных библиотек.

Литература

[1] О. М. Атаева, В. А. Серебряков, Подход к созданию персональной электронной семантической библиотеки, RCDL, 2013.

[2] Bernhard Haslhofer, Bernhard Schan, The OAI2LOD Server: Exposing OAI-PMH Metadata as Linked Data, 2008. http://eprints.cs.univie.ac.at/284/1/lodws2008.pdf

[3] http://dbpedia.org

[4] http://europeana.eu

[5] R. Weber. Ontological Foundations of Information Systems, Queensland, Australia, Coopers & Lybrand. 1997.

[6] О. М. Атаева, А. О. Еркимбаев, В. Ю. Зицерман, Г. А. Кобзев, К. П. Пушин, В. А. Серебряков, К. Б. Теймуразов. Интеграция данных по теплофизическим свойствам веществ методами онтологического моделирования, RCDL, 2013.

[7] http://www.w3.org/TR/void/

[8] P. Jain, K. Verma, P.Z. Yeh, P. Hitzler, A.P. Sheth. LOQUS: Linked Open Data SPARQL Querying System. Technical report, Tech. rep., Kno. e. sis Center, Wright State University, Dayton, Ohio, 2010. Available from http://www.pascal-hitzler.de/resources/publications/loqus-tr-2010.pdf

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

[10] А. Антопольский, О. Атаева, В. Серебряков. Среда интеграции данных научных библиотек, архивов и музеев «LibMeta» // Информационные ресурсы России. – 2012. – № 5. – С. 8–12.

[11] http://e-heritage.ru/index.html

[12] http://lod2.eu/Project/Silk.html

[13] http://viaf.org/

[14] http://pro.europeana.eu/ese-documentation/

[15] http://d2rq.org/d2r-server

[16] Е. Горный. Развитие электронных библиотек: мировой и российский опыт, проблемы, перспективы / Е. Горный, К. Вигурский // Интернет и российское общество / под ред. И. Семенова; Моск. Центр Карнеги. – М. : Гендальф, 2002. – С.158–188.

Personal Digital Library Libmeta as an Integrating Environment for Linked Open Data

Olga M. Ataeva, Vladimir A. Serebryakov

The article describes semantic digital library Libmeta resources of which can be enriched by means of using data from the sources located in LOD. Binding is due to domain ontology which is user defined and determines his/her field of interest. Problems of integration of library resources in LOD and creation of search queries on data sources are considered as well as use of specifications and technologies from LOD stack within a system considered.