СЕМАНТИЧЕСКАЯ ИНТЕГРАЦИЯ БИБЛИОТЕЧНЫХ ДАННЫХ

В.А.Серебряков[a], *К.Б.Теймуразов*[a], *О.Н.Шорин*[b]

[a] Вычислительный центр им.А.А.Дородницына РАН, Москва
[b] Российская национальная библиотека, Санкт-Петербург

В Российской государственной библиотеке и Российской национальной библиотеке начат совместный проект, целью которого является публикация библиотечных данных библиотек, входящих в состав Национальной электронной библиотеки, в соответствии с принципами Linked Open Data. Реализация данного проекта позволит получить доступ к библиографической информации, хранящейся в ряде крупнейших библиотек России, в виде, пригодном для машинной обработки. Набор данных состоит из нескольких десятков миллионов записей. В процессе семантической интеграции предстоит решить ряд актуальных задач: разработка онтологии предметной области, конвертация библиотечных данных из различных MARC-форматов в RDF, публикация данных и предоставление SPARQL точек доступа к ним. Наличие открытого доступа к одному из самых крупных в мире массиву библиографической информации с возможностью обнаружения семантически связанных данныхбудет являться одной из составляющих развития как культуры в целом, так и отдельных направлений книжной отрасли в частности.

Ключевые слова: Открытые связанные данные, семантическая паутина, связанные данные, библиографическая запись.

The Russian State Library and Russian National Library launched a joint project, the aim of which is to publish the library in accordance with the principles of Linked Open Data. This project will provide access to bibliographic information stored in a number of the largest libraries of Russia, in a form suitable for machine processing. The data set consists of several tens of millions of records. A number of pressing problems will be solved in the process of semantic integration: the development of the domain ontology, the conversion of library data from various MARC-formats to RDF, the publication of data and the provision of SPARQL access points to them.

Keywords: Linked Open Data, semantic Web, bibliographic record.

В большинстве своем, информация, представленная на сайтах, предназначается для людей, поскольку основу интернета составляет гипертекст. Это означает, что основный смысл, значение скрывается в самом тексте, что значительно усложняет процесс извлечения этой сути, пригодного для автоматизированной обработки. В 2006 году Тимом Бернерсом-Ли была предложена надстройка над существовавшим интернетом, которая позволила бы автоматизированным системам извлекать информацию, анализировать её, устанавливать взаимосвязи и генерировать новую информацию. Такой подход он назвал «семантической паутиной».

Тим Бернерс-Ли предложил использовать термин «связанные данные» для реализации семантической паутины. Основное отличие семантической паутины заключается именно в термине «данные», которое ставилось в противовес существовавшему на тот момент, пусть и «гипер-», но всё же «тексту». В нашей жизни мы оперируем множеством данных: информация о стоимости продуктов в магазине, расписание авиарейсов, информация об авторстве литературного произведения.

Анализируя данные, человек может принять взвешенное решение. Например, имея данные о наличии и стоимости книги в разных книжных магазинах, а также информации о месторасположении и часах работы этих магазинов, человек способен сделать выбор и купить необходимую ему книгу по оптимальной цене в близлежащем работающем магазине. К сожалению, автоматизировать этот процесс в терминах гипертекста чрезвычайно сложно [1].

Для оперирования данными необходимо было решить несколько ключевых вопросов [2]:

- Каким образом обеспечить доступ к данным, чтобы их можно было повторно использовать?

- Как должно происходить обнаружение данных, связанных с уже имеющимися данными?
- Как приложения должны интегрировать разнородные данные, полученные из большого числа заранее неопределенных источников?

Также как World Wide Webизменил способы работы с текстом, с документами, необходимо было придумать механизмы поиска, доступа, интеграции и использования данных.

Тим Бернерс-Ли сформулировал четыре основных принципа связанных данных [3]:

1. Применение универсальных идентификаторовURIв качестве имен сущностей.
2. ПрименениеHTTPURIдля реализации возможности обращения по именам, чтобы они могли быть найдены как людьми, так и программными системами.
3. Предоставление полезной информации о сущности при обращении по URI, используя стандартизованные форматы.
4. Включение ссылок на другие связанные URI для облегчения поиска.

Для реализации этих принципов было предложено использовать модель преставления данных RDF (Resource Description Framework), которая пригодна для машинной обработки. Структурно выражения в RDFпредставляют собой триплеты. Каждый триплет состоит из субъекта, предиката и объекта. Выражение RDF-триплета означает, что отношение, указанное предикатом, связывает предметы, обозначенные как субъект и объект [4]. Например, предикат «является автором» может связывать субъект «Достоевский» и объект «Преступление и наказание». Основная

идея RDFсостоит в том, чтобы показать взаимосвязь одних данных с другими.

RDF не является форматом, а представляет собой абстрактную модель для описания взаимоотношений между данными в виде триплетов. Для сериализации RDF-триплетов существует несколько способов. Наиболее распространенным способом является представление в виде XML -RDF/XML. Синтаксис RDF/XMLстандартизован консорциумом W3C и широко используется для публикации связанных данных в интернете.

Для встраивания RDF-триплетов непосредственно в HTML-документы используют формат сериализации RDFa. Изначально RDFинформацию указывали в виде комментариев в HTML-документах, однако впоследствии триплеты стали органично встраивать в объектную модель документа (Document Object Model, DOM).

Существует способ сериализации RDF, ориентированный на создание и чтение триплетов человеком –Turtle. N-Triplesявляется подмножеством Turtle, в котором отсутствует возможность использования пространства имен (namespaces) и других методов сокращенияразмера файла, например, компактные URI (CURIE) или вложенные конструкции. В связи с этим файл, написанный с использованием N-Triples, получается гораздо больше, чем с использованием Turtlеи даже RDF/XML. Но у N-Triplesесть одно неоспоримое преимущество: благодаря отсутствию механизмов сокращения размера файла каждая строка содержит в себе исчерпывающий объем информации, поэтому файл N-Triplesможет быть считан и разобран построчно.

Множество современных языков программирования поддерживают нотацию JSON, поэтому неудивительно, что существует способ сериализации RDF/JSON.

RDF представляет собой абстрактную модель представления данных с помощью триплетов и никоим образом не затрагивает семантики описываемых данных. Для выражения семантики используются словари, таксономии и онтологии, которые задаются с использованием языков RDFS (RDF Vocabulary Description Language), SKOS (Simple Knowledge Organization System) и OWL (Web Ontology Language) соответственно [5].

SKOSпредставляет собой словарь иерархически организованных терминов, а RDFS и OWL являются словарями для описания концептуальных свойств в терминах классов, свойств, экземпляров классов и операций. Например, формальная семантика OWLописывает, как получать логические следствия, т.е. факты, которые не представлены в онтологии буквально, но следуют из ее семантики.

С использованием принципов, предложенных Тимом Бернерсом-Ли, в интернете реализуется проект открытых связанных данных (Linked Open Data), целью которого является интеграция данных, информации и знаний посредством глобальных идентификаторов ресурсовURIи моделью данных RDF.

Библиотеки нашей страны хранят у себя множество различных данных: информация о записавшихся в библиотеку читателях, имеющихся в наличии книг, отсканированных образах различных изданий. Среди множества хранящихся в библиотеках данных особое значение имеет библиографическая информация, выраженная в виде библиографических записей, создаваемых непосредственно в библиотеках в процессе каталогизации книг.

«Библиографическая запись - элемент библиографической информации, фиксирующий в документальной форме сведения о документе – объекте записи, позволяющие его идентифицировать, раскрыть его состав и содержание в целях библиографического поиска. В состав библиографической записи входит биб-

лиографическое описание, дополняемое, по мере необходимости, заголовком, терминами индексирования (классификационными индексами и предметными рубриками), аннотацией (рефератом), шифром хранения документа, дополнительными точками доступа, сведениями о связи с другими библиографическими записями и другой дополнительной информацией о документе, обеспечивающей доступ к нему, датой завершения обработки документа, сведениями служебного характера» [6].

С точки зрения связанных данных библиографические записи представляют огромный интерес, поскольку хранящаяся в них информация взаимосвязана: авторы связаны со своими произведениями, сериальные издания связаны друг с другом через общую часть, издательства имеют непосредственное отношение к изданным у них книгам и т.д.

В мировом сообществе реализуется ряд проектов, направленных на публикацию библиографической информации в Linked Open Data. В частности, одним из первых проектов этом направлении являлась инициатива Библиотеки Конгресса (Library of Congress), в рамках которой было опубликовано более 260 тысяч авторитетных записей. Стоит отметить также проект создания Виртуального Международного Авторитетного Файла (Virtual International Authority File), в котором участвуют более 35 национальных библиотек мира [7]. Целью этого проекта является сопоставление одних и тех же авторитетных записей из разных библиотек мира.

Проект The Open Library можно смело назвать наиболее амбициозным, поскольку его конечной целью является создание отдельной веб-страницы для каждой выпущенной книги. На данный момент на сайте представлена информация о 20 миллионах книг и 6 миллионах авторов.

В Министерстве культуры Российской Федерации предпринимаются попытки, направленные на реализацию нового этапа развития Национальной электронной библиотеки (НЭБ). Основной целью этого этапа является обеспечение свободного, равного и всеобщего доступа граждан нашей страны к документной информации историко-культурного, научного и образовательного назначения через сеть Интернет, предоставляемой на основе единой общенациональной системы создания и эффективного использования цифровых библиотечно-информационных ресурсов и сервисов [8].

Достижение поставленной цели будет осуществляться путем решения ряда задач:

1. Формирование распределенного фонда, в состав которого будут входить актуальные научные и образовательные материалы, востребованные жителями страны произведения, социально значимая информация.
2. Обеспечение доступа к распределенному цифровому фонду путем создания единой точки доступа, предоставляющей развитый набор сервисов по поиску материалов в распределенном массиве информации.
3. Решение нормативно-правовых аспектов деятельности Национальной электронной библиотеки, в частности унификация содержания государственных заданий для различного вида библиотек с возможностью внесения изменений в перечни оказываемых электронных услуг.

В процессе реализации нового этапа развития НЭБ из библиотек различной ведомственной подчиненности будет аккумулирована уникальная по своей полноте библиографическая информация. Публикация собранных данных в семантически связанном виде выведет НЭБ в ряды лидеров проектов в мировом библио-

течном сообществе, как по объемам опубликованных данных, так и по количеству источников, участвующих в интеграции.

В Российской государственной библиотеке и Российской национальной библиотеке был реализован совместный проект, целью которого являлосьсоздание программной системы, позволяющей осуществить публикацию данных библиотек, входящих в состав НЭБ, в соответствии с принципами Linked Open Data. Архитектура данной программной системы предусматривает инфраструктурные особенности функционирования НЭБ, в частности децентрализацию процессов формирования, хранения фондов и вариативность технологических решений, используемых в отдельно взятых библиотеках – участниках НЭБ.

В рамках реализации программной системыбыл решен ряд принципиальных задач.

1. Разработка онтологии предметной области на базе существующих решений. При создании онтологии предметной области максимально использовались термины из уже существующих и широко используемых словарей [9]. Такой подход значительно снизил вероятность того, что для существующих программных систем могла потребоваться дополнительная конвертация данных или даже изменение приложения.

Были изучены проекты Библиотеки конгресса США, прежде всего стандарт METSпредставления описательных, административных и структурных метаданных цифровых библиотек. Также был исследован проект Europeana, который в качестве метаданных использует стандарт Dublin Core [10]. Немаловажным было изучение опыта проекта Delos, который выпустил документ Digital Library Reference Model. Также был учтен стандарт Publishing Requirements for Industry Standard Metadata (PRISM), разработанный издательствами для обмена метаданными о публикациях.

2. Осуществление интеграции с автоматизированными библиотечными информационными системами (АБИС) и конвертация библиографических записей в унифицированный формат. Для автоматизации процесса комплектования, каталогизации, книговыдачи, межбиблиотечного обмена большинство библиотек используют АБИС. Как следствие, все библиографические описания, имеющиеся в библиотеке, хранятся в АБИС. Библиотеки, являющиеся участниками и партнерами Национальной электронной библиотеки, используют 4 основных АБИС: Aleph, Ирбис, MarcSQL, Opac-Global.

Перечисленные АБИС имеют широкие возможности по интеграции с внешними системами, используя различные протоколы. Для каждой АБИС были изучены различные возможности подключения, позволяющие экспортировать библиографические описания. В частности, для Aleph модуль интеграции был реализован с помощью использования протокола OAI-PMH [11], а для системы Ирбис запрос библиографического описания осуществлялся по протоколу Search/Retrieve via URL версии 1.1. Интеграция с MarcSQL осуществляется путем прямого доступа к базе данных АБИС. Из Opac-Global описания экспортируются путем прямых HTTP-запросов к служебным адресам системы.

Записи, получаемые из АБИС библиотек, представлены в MARC-форматах. В России используются 2 различных формата хранения библиографических описаний: MARC21 и RUSMARC (диалект формата UNIMARC). Оба этих формата являются бинарными. MARC21 – это международный формат, разработанный Библиотекой Конгресса. Для этого формата существует множество утилит, позволяющих конвертировать файлы в MARC21/XML. В конечном итоге записи, полученные в формате MARC21, конвертируются в формат MODS [12]. В силу малой распространенности формат RUSMARC не имеет утилит по конвертации из бинарного вида в представление, основанное на использовании XML. В связи с этим был разработан конвертор, который преобразует записи из формата RUSMARC в MODS.Для экспортиро-

ванных записей было создано единое хранилище, которое ежедневно пополняется новыми и отредактированными описаниями в формате MODS из АБИС библиотек.

3. Осуществление взаимного обогащения данных из различных библиотек. В случае появления в хранилище нескольких библиографических записей на одну и ту же книгу или авторитетных записей на одного и того же автора из различных библиотек, отличающихся друг от друга по степени детализации, раскрытия информации, наличием точек доступа, ссылок, автоматически создается или уточняетсяобъединенная запись, максимально полно раскрывающая первоисточники.

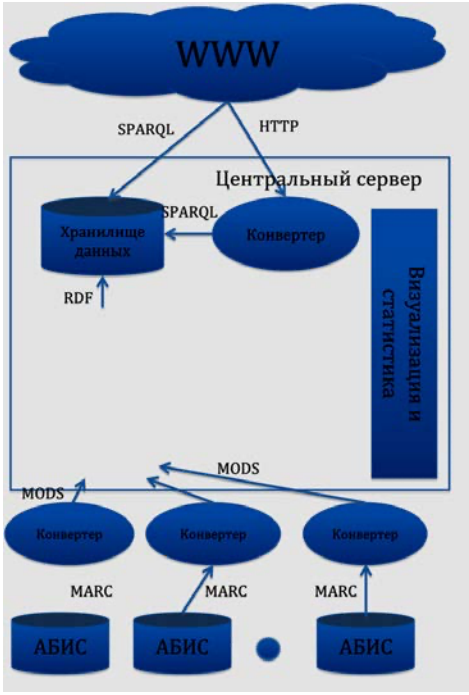

Рис. 1. Архитектура системы интеграции.

4. Конвертация и решение вопроса о хранении данных. К обогащенным записям, расположенным в едином хранилище, применяется преобразование, которое трансформирует записи из формата MODS в формат RDF, в соответствии с предметной онтологией. Для обеспечения точки доступа к RDFданным с помощью языка запросов SPARQL (SPARQL Protocol and RDF Query Language) был разработан механизм хранения RDF-триплетов. Для этого были проанализированы несколько подходов: автоматическая конвертация MODS в RDF-триплеты «на лету» для каждого запроса, хранение заранее сконвертированных данных в реляционной базе данных, хранение данных в специализированном хранилище триплетов. Каждый из этих подходов имеет свои преимущества и недостатки [13].

Например, автоматическая конвертация данных по каждому запросу не приводит к дублированию данных, но требует реализации сложной логики и обладает низкой производительностью. Хранение же триплетов, в свою очередь, приводит к дублированию данных, что требует дополнительного физического пространства и механизмов модификации триплетов в случае изменения библиографических записей. Поскольку объем информации, хранимой в библиографических записях, является несущественным, а изменение информации происходит только в одном направлении, что устраняет вероятность возникновения коллизий, было принято решение о хранении данных в специализированном хранилище триплетов.

5. Выбор данных для связывания и публикация записей в Linked Open Data. По правилам публикации данных в LOD новые сущности должны ссылаться на уже опубликованные наборы. Для этого были исследованы уже опубликованные массивы данных на предмет возможности использования их в качестве субъектов в RDF-триплетах [14].Использование специализиро-

ванного хранилища триплетов позволило автоматически создать SPARQLточку доступа к данным и обертки вокруг неё в виде обычного веб-сервера.

6. Реализация модуля визуализации полученного результата. Для отладки всего процесса публикации обогащенных записей в Linked Open Data и верификации результата был создан веб-сайт, на котором визуально отображены исходные записи, полученные из различных источников, обогащенная запись, результаты, опубликованные в LOD. Интерфейс позволяет строить отчеты на основе статистической информации и экспортировать их в различных форматах: xls, csv, xml, html.

Положительный эффект от публикации библиотечных данных в семантически связанном виде, пригодном для машинного использования, трудно переоценить. В процессе реализации этого проекта был решен ряд принципиальных задач, связанных с разнородностью используемых российскими библиотеками программных систем, форматов представления данных, протоколов взаимодействия. Для достижения поставленной целибыл использован опыт передовых библиотек мира, адаптированный к специфике каталогизации литературы в России. В результате была создана модульная система, способная с использованием минимальных усилий подключать новые библиотеки в качестве источников библиографических данных.

Список литературы

1. Berners-Lee T., James H., Lassila O. The Semantic Web. Scientific American Magazine, March 26, 2008.
2. Heath T., Bizer C. *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool.

3. Berners-Lee T. Linked Data – Design Issues. http://www.w3.org/DesignIssues/LinkedData.html
4. Спецификация языка RDF. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/
5. Спецификация языка OWL. http://www.w3.org/TR/2012/REC-owl2-syntax-20121211/
6. Российский коммуникативный формат представления библиографических записей в машиночитаемой форме. http://www.rusmarc.ru/rusmarc/format.html
7. VIAF project description. http://www.oclc.org/viaf.en.html
8. Заседание коллегии Министерства культуры РФ от 23.04.2014г. http://mkrf.ru/m/494838/
9. Hannemann J., Kett J. Linked Data for Libraries. 76TH IFLA GENERAL CONFERENCE, 2010.
10. Haslhofer B., Isaac A. data.europeana.eu. The Europeana Linked Open Data Pilot. Proc. Int'l Conf. on Dublin Core and Metadata Applications, 2011.
11. The Open Archives Initiative Protocol for Metadata Harvesting. http://www.openarchives.org/OAI/openarchivesprotocol.html
12. Metadata Object Description Schema. http://www.loc.gov/standards/mods/mods-overview.html
13. Böhme C. Towards an Infrastructure for the Synchronisation of Library Metadata. Semantic Web in Libraries, 2012.
14. Volz J., Bizer C., Gaedke M., Kobilarov G. Discovering and maintaining links on the web of data. In Proceedings of the International Semantic Web Conference, pages 650–665, 2009.