УДК 025.34:004.9
ББК 78.362.5

Проблемы семантической интеграции библиотечных данных

В статье рассказывается о совместном проекте Российской государственной библиотеки и Российской национальной библиотеки. Его цель — публикация данных участников Национальной электронной библиотеки в соответствии с принципами Linked Open Data. Реализация проекта позволит получить доступ к библиографической информации, хранящейся в ряде крупнейших библиотек России, в виде, пригодном для машинной обработки. Перечислены принципиальные задачи, которые предстоит решить в процессе семантической интеграции библиотечных данных.

Ключевые слова: *Linked Open Data, семантическая паутина, связанные данные, библиографическая запись.*

Владимир Алексеевич Серебряков,
заведующий отделом Вычислительного центра им. А.А. Дородницына Российской академии наук, доктор физико-математических наук (Москва)

Олег Николаевич Шорин,
заместитель генерального директора по информатизации Российской национальной библиотеки (Санкт-Петербург)

«Отцом» современного Интернета считается Тим Бернерс-Ли, хотя, строго говоря, это не совсем так. Протоколы, с помощью которых компьютеры обменивались данными, были созданы в Агентстве по перспективным оборонным исследовательским проектам (DARPA) Министерства обороны США в конце 1960-х годов.

В 1989 г. Т. Бернерс-Ли предложил мировому сообществу концепцию Всемирной паутины. Работая в Европейском центре ядерных исследований (CERN), он обратил внимание на разрозненность информации, хранившейся на разных компьютерах. Процесс передачи информации с одного компьютера на другой нередко сопровождался многочисленными сложностями, связанными как с несовместимостью форматов, так и наличием иерархической структуры подчинения отделов. Т. Бернерс-Ли предложил использовать принцип перекрестных ссылок для того, чтобы исследователь мог получить информацию, находящуюся на другом компьютере, непосредственно с него, не копируя ее и не проходя все иерархическое дерево подчинения отделов. Проблему различий в операционных системах он предложил решить с помощью установки клиентского программного

обеспечения, которое могло бы переходить по ссылкам, размещенным в гипертексте.

В течение нескольких следующих лет он разработал протокол передачи гипертекста HTTP, язык гипертекстовой разметки HTML, глобальные идентификаторы ресурсов URI. Предложенные им технологии используются и по сей день. Основная заслуга Т. Бернерса-Ли заключается в том, что он создал единый стандарт, который позволил Интернету остаться неделимым, не распавшись на непересекающиеся друг с другом множества частных сетей. Он же придумал термин «Всемирная паутина» (World Wide Web).

Изначально Интернет представлял собой среду, в которой общались технические специалисты, ученые из различных областей науки. Для создания и наполнения сайта требовались глубокие познания в программировании на различных языках, в том числе HTML. На этом этапе контент создавался узким кругом высококвалифицированных специалистов и был предназначен для потребления профессионалами.

В конце XX в. произошел знаменитый «кризис доткомов», в результате чего было несколько подорвано доверие к Интернету. Через несколько лет после этого появился термин Web 2.0, основная суть которого состояла в том, что контент создавался не группой избранных лиц, а всеми участниками Интернета. Появились инструменты, упрощавшие ввод, загрузку информации на сайты (блоги, комментарии, социальные сети). Существует мнение, что именно для реабилитации всемирной паутины был предложен термин Web 2.0, который заменил устаревший Web 1.0. Однако большинство ученых, в том числе Т. Бернерс-Ли, считают, что Web 2.0 — это всего лишь удачный маркетинговый ход, а не новая технология, поскольку базируется на тех же протоколах, которые использовались с самого начала существования Интернета.

В любом случае информация, представленная на сайтах, предназначалась для людей, поскольку основу Интернета составлял гипертекст. Основный смысл скрывался в самом тексте, и не существовало формального способа извлечения информации, пригодной для автоматизированной обработки. Т. Бернерс-Ли предложил надстройку над существовавшим Интернетом, которая позволила бы автоматизированным системам извлекать информацию, анализировать ее, устанавливать взаимосвязи и генерировать новую информацию. Такой подход он назвал «семантическая паутина».

Связанные данные

Т. Бернерс-Ли ввел термин «связанные данные» (Linked Data) для реализации семантической паутины. Основное отличие семантической паутины заключается именно в термине «данные» (в противовес существовавшему на тот момент пусть и «гипер-», но все же «тексту»). Ежедневно человек оперирует множеством данных: информацией о стоимости продуктов в магазине, об авторстве литературного произведения, о расписании авиарейсов и др. Анализируя их, можно принять взвешенное решение. Например, имея данные о наличии и стоимости книги в разных книжных магазинах, а также информацию о расположении 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 [4].

Для оперирования данными необходимо было решить несколько ключевых вопросов:

- каким образом обеспечить доступ к данным, для того 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 [7].

Необходимо было придумать механизмы поиска, доступа, интеграции и использования данных. В 2006 г. Т. Бернерс-Ли сформулировал четыре основных принципа связанных данных:

• применение универсальных идентификаторов URI в качестве имен сущностей;

• применение HTTP URI для реализации возможности обращения по именам, для того чтобы они могли быть найдены как людьми, так и программными системами;

• предоставление полезной информации о сущности при обращении по URI, используя стандартизованные форматы;

• включение ссылок на другие связанные URI для облегчения поиска [3].

С целью реализации этих принципов было предложено использовать модель представления данных RDF (Resource Description Framework), пригодную для машинной обработки. Структурно выражения в RDF являются триплетами. Каждый триплет состоит из субъекта, предиката и объекта. Выражение RDF-триплета означает, что отношение, указанное предикатом, связывает предметы, обозначенные как субъект и объект [10]. Например, предикат «является автором» может связывать субъект «Достоевский» и объект «Преступление и наказание». Основная идея RDF состоит в том, чтобы показать взаимосвязь одних данных с другими.

RDF не является форматом, это абстрактная модель для описания взаимоотношений между данными в виде триплетов. Для сериализации RDF-триплетов существует несколько способов. Наиболее распространенным способом является представление в виде XML. Синтаксис RDF/XML стандартизован Консорциумом Всемирной паутины (W3C) и широко используется для публикации связанных данных в Интернете.

Для встраивания RDF-триплетов непосредственно в документы HTML используют формат сериализации RDF. Изначально RDF-информацию указывали в виде комментариев в документах HTML, 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. Подмножеством Turtle является N-Triples, в нем отсутствует возможность использования пространства имен (namespaces) и других методов сокращения размера файла, например, компактных URI (CURIE) или вложенных конструкций. Поэтому файл, написанный с использованием N-Triples, получается гораздо больше, чем с использованием Turtle и даже RDF/XML. Но у N-Triples есть одно неоспоримое преимущество: благодаря отсутствию механизмов сокращения размера файла каждая строка содержит в себе исчерпывающий объем информации, и файл N-Triples может быть считан и разобран построчно.

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

RDF никоим образом не затрагивает семантику 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 [9].

SKOS представляет 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 OWL описывает, как получать логические следствия, т. е. факты, которые не представлены в онтологии буквально, но следуют из ее семантики.

С использованием принципов, предложенных Т. Бернерсом-Ли, в Интернете реализуется проект Linked Open Data (LOD, открытые связанные данные), целью которого является интеграция данных, информации и знаний посредством глобальных идентификаторов ресурсов URI и моделью данных RDF.

Семантическая интеграция библиотечных данных

Библиотеки хранят у себя множество различных данных: информацию о читателях, имеющихся в наличии книг, отсканированных образах различных изданий. Среди множества 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 библиографическое 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 до-

ступ к нему, датой завершения обработки документа, сведениями служебного характера» [2].

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 ими книгам и т. д.

В мировом сообществе реализуется ряд проектов, направленных на публикацию библиографической информации в LOD. В частности, одним из первых проектов в этом направлении являлась инициатива Библиотеки Конгресса США, в рамках которой было опубликовано более 260 тыс. авторитетных записей. Следует отметить также проект создания Виртуального международного авторитетного файла (Virtual International Authority File), в котором участвуют более 35 национальных библиотек [11]. Целью проекта является сопоставление одних и тех же авторитетных записей из различных библиотек мира.

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

Постановка задачи

В Министерстве культуры Российской Федерации предпринимаются попытки, направленные на реализацию нового этапа развития Национальной электронной библиотеки (НЭБ). Основная цель этого этапа — обеспечение свободного, равного и всеобщего доступа граждан России к документной информации историко-культурного, научного и образовательного назначения через Интернет, предоставляемой на основе единой общенациональной системы создания и эффективного использования цифровых библиотечно-информационных ресурсов и сервисов [1].

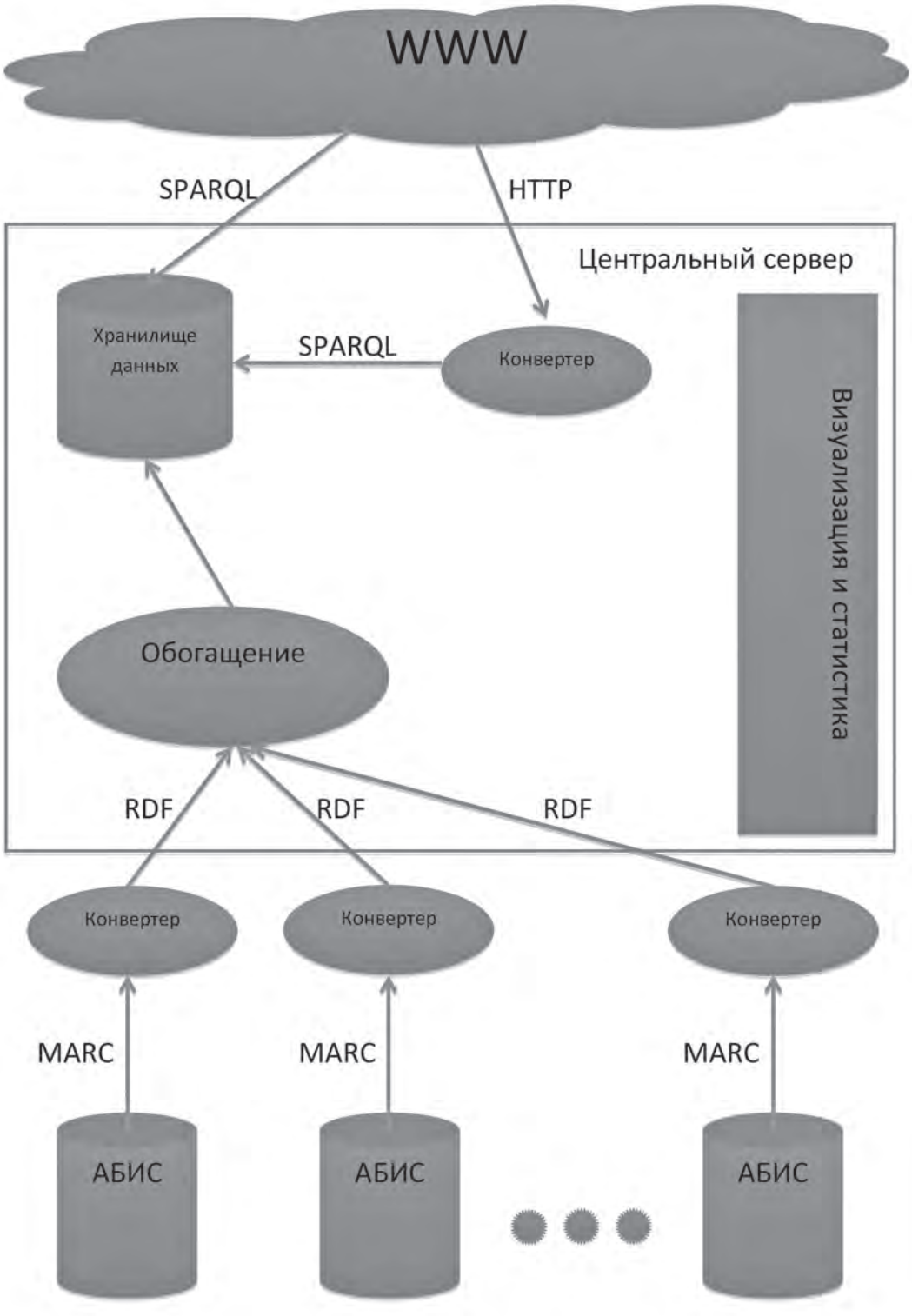
Для достижения поставленной цели необходимо решить следующие задачи:

- формирование распределенного фонда, в состав которого будут входить актуальные научные и образовательные материалы, востребованные жителями страны произведения, социально значимая информация;
- обеспечение доступа к распределенному цифровому фонду путем создания единой точки доступа, предоставляющей развитый набор сервисов по поиску материалов в распределенном массиве информации;
- урегулирование нормативно-правовых аспектов деятельности НЭБ, в частности унификация содержания государственных заданий для различных видов библиотек с возможностью внесения изменений в перечень оказываемых электронных услуг.

В процессе реализации нового этапа развития НЭБ из библиотек разной 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 библиотечном сообществе как по объемам опубликованных данных, так и по количеству источников, участвующих в интеграции.

Архитектура системы интеграции

В Российской государственной библиотеке и Российской национальной библиотеке начат совместный проект, целью которого является создание программной системы, позволяющей осуществить публикацию данных библиотек, входящих в состав НЭБ, в соответствии с принципами LOD. Архитектура программной системы должна предусматривать инфраструктурные особенности функционирования НЭБ, в частности



Архитектура системы интеграции

децентрализацию процессов формирования, хранения фондов и вариативность технологических решений, используемых в отдельно взятых библиотеках — участниках НЭБ.

В рамках реализации программной системы должны быть решены некоторые принципиальные задачи.

1. *Разработка онтологии предметной области на базе существующих решений.*

При создании онтологии предметной области необходимо максимально использовать термины из широко используемых словарей [6]. Такой подход значительно снижает вероятность того, что для существующих программных систем может потребоваться дополнительная конвертация данных или даже изменение приложения. Следует изучить проекты Библиотеки Конгресса США, прежде всего стандарт METS представления описательных, административных и структурных метаданных цифровых библиотек, а также проект Europeana, который в качестве метаданных использует стандарт Dublin Core [8]. Немаловажным будет изучение опыта проекта Delos и документа Digital Library Reference Model. Необходимо также учитывать стандарт PRISM (Publishing Requirements for Industry Standard Metadata), разработанный издательствами для обмена метаданными о публикациях.

2. *Осуществление интеграции с автоматизированными библиотечными информационными системами (АБИС).*

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 НЭБ, используют в основном четыре АБИС: Aleph, Ирбис, MarcSQL, Opac-Global, имеющие широкие возможности по интеграции с внешними системами с использованием различных протоколов. Для каждой АБИС необходимо изучить различные возможности подключения, выбрать способ, который удовлетворяет всем потребностям, и реализовать программный модуль взаимодействия.

3. *Осуществление конвертации библиографических записей в унифицированный формат.*

В России используются два различных формата хранения библиографических описаний — MARC21 и RUSMARC (диалект формата UNIMARC). Оба формата являются бинарными. MARC21 — это международный формат, разработанный Библиотекой Конгресса США. Для него существует множество утилит, позволяющих конвертировать файлы в MARC21/XML. В силу малой распространенности формат RUSMARC не имеет утилит по конвертации из бинарного вида в представление, основанное на использовании XML. Необходимо сконвертировать полученные из библиотек данные в формат RDF, согласованный с предметной онтологией.

4. *Разработка механизма хранения сконвертированных данных.*

Для обеспечения точки доступа к RDF-данным с помощью языка запросов SPARQL необходимо разработать механизм хранения RDF-триплетов. Должны быть проанализированы несколько подходов: автоматическая конвертация MARC-данных в RDF-триплеты «на лету» для каждого запроса, хранение заранее сконвертированных данных в реляционной базе данных, хранение данных в специализированном хранилище триплетов. Каждый из подходов имеет свои преимущества и недостатки [5]. Например, автоматическая конвертация данных по каждому запросу не приводит к их дублированию, но требует реализации сложной логики и будет обладать низкой производительностью. Хранение же триплетов, в свою очередь, является причиной дублирования данных. Это потребует дополнительного физического пространства и механизмов модификации триплетов в случае изменения библиографических записей.

5. *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 степенью детализации, раскрытия информации, наличием точек доступа, ссылок, должна быть создана или обогащена объединенная запись, максимально полно раскрывающая первоисточники.

6. *Выбор данных для связывания и публикация записей в LOD.*

По правилам публикации данных в LOD новые сущности должны ссылаться на уже опубликованные наборы. Для этого необходимо будет исследовать уже опубликованные массивы данных на предмет возможности использования их в качестве субъектов в RDF-триплетах [12]. Следует провести анализ имеющихся механизмов публикации данных в LOD, выбрать наиболее подходящие для поставленной задачи и осуществить публикацию с их использованием. Необходимо будет создать также точку доступа SPARQL к данным и обертки вокруг нее в виде обычного веб-сервера.

7. *Реализация модуля визуализации полученного результата.*

Для отладки всего процесса публикации обогащенных записей в LOD понадобится механизм верификации результата. Нагляднее всего это делать с помощью веб-сайта, на котором визуально были бы отображены исходные записи, полученные из различных источников, и обогащенная запись, опубликованная в LOD.

Положительный эффект от публикации библиотечных данных в семантически связанном виде, пригодном для машинного использования, трудно переоценить. Однако в процессе реализации этого проекта необходимо решить ряд принципиальных задач, связанных с разнородностью используемых российскими библиотеками программных систем, форматов представления данных, протоколов взаимодействия. Для достижения поставленной цели следует использовать опыт передовых библиотек мира, адаптируя его к специфике каталогизации литературы в России. В результате будет создана модульная система, способная при минимальных усилиях подключать новые библиотеки в качестве источников библиографических данных.

Список источников

1. Новости Министерства // М-во культуры Рос. Федерации : Официальный сайт [Электронный ресурс]. — Режим доступа: http://mkrf.ru/m/494838/
2. Российский коммуникативный формат представления библиографических записей в машиночитаемой форме (российская версия UNIMARC) [Электронный ресурс]. — Режим доступа: http://www.rusmarc.ru/rusmarc/format.html
3. *Berners-Lee T.* Linked Data [Electronic resource]. — Mode of access: http://www.w3.org/DesignIssues/LinkedData.html
4. Idem. The Semantic Web [Electronic resource] / T. Berners-Lee, J. Hendler, O. Lassila // Scientific American Magazine. — Mode of access: http://www.cs.umd.edu/~golbeck/LBSC690/SemanticWeb.html
5. *Böhme C.* Towards an Infrastructure for the Synchronisation of Metadata in Library [Electronic resource] // SNIB12 Semantic Web in Libraries. — Mode of access: http://swib.org/swib12/programme.php
6. *Hannemann J.* Linked Data for Libraries [Electronic resource] / J. Hannemann, J. Kett // World Library and Information Congress: 76th IFLA General Conference and Assembly. — Mode of access: http://conference.ifla.org/past-wlic/2010/149-hannemann-en.pdf
7. *Heath T.* Linked Data: Evolving the Web into a Global Data Space : Synthesis Lectures on the Semantic Web: Theory and Technology [Electronic resource] / T. Heath, C. Bizer. — Mode of access: http://www.morganclaypool.com/doi/abs/10.2200/S00334ED1V01Y201102WBE001
8. I*saac A.* Europeana Linked Open Data — data.europeana.eu [Electronic resource] / A. Isaac, B. Haslhofer. — Mode of access: http://www.semantic-web-journal.net/system/files/swj297_1.pdf
9. OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax (Second Edition) [Electronic resource]. — Mode of access: http://www.w3.org/TR/2012/REC-owl2-syntax-20121211/
10. Resource Description Framework (RDF) : Concepts and Abstract Syntax [Electronic resource]. — Mode of access: http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/
11. VIAF. Virtual International Authority File [Electronic resource]. — Mode of access: http://www.oclc.org/viaf.en.html
12. *Volz J.* Discovering and Maintaining Links on the Web of Data / J. Volz [et al.] // In Proceedings of the International Semantic Web Conference. — Chantilly (VA, USA), 2009. — P. 650—665.

Контактные данные:
e-mail: serebr@ultimeta.ru, shorin@nlr.ru