О. ШОРИН

УДК 004.4

ПУБЛИКАЦИЯ БИБЛИОГРАФИЧЕСКИХ ДАННЫХ В ОТКРЫТОЕ СЕМАНТИЧЕСКОЕ ПРОСТРАНСТВО

Введение

Для процесса каталогизации в библиотеках используются автоматизированные библиотечные интегрированные системы (АБИС), которые позволяют реализовать различные технологические цепочки создания и редактирования библиографических записей. В каждой отдельно взятой библиотеке приняты собственные правила обработки и описания книг, которые зависят как от каналов поступления экземпляров, например, обязательный экземпляр, поставляемый Книжной палатой, и дар автора будут обработаны различными способами, так и от вида литературы - периодические издания, такие как газеты, журналы, обрабатываются несколько иначе по сравнению с книгами. Часть экземпляров может отправляться на выставку новых поступлений или попадать в обменный фонд. Количество различных путей обработки экземпляров может варьироваться в каждой библиотеке в зависимости от размера учреждения, профиля его комплектования, внутренних технологических особенностей.

После того, как экземпляр произведения проходит полный цикл обработки и попадает на полку, в АБИС сохраняется библиографическая запись, относящаяся к этому экземпляру. В своей базе данных АБИС хранит библиографическую запись в некоем внутреннем формате, однако для взаимодействия с внешними системами используется коммуникативный формат представления библиографической записи в машиночитаемой форме. Наибольшее распространение получили форматы семейства MARC (Machine-Readable Cataloging). К плюсам формата MARC можно отнести:

- широкую распространенность формата;
- наличие множества полей, позволяющих детально описать объект.

Российской версией коммуникативного формата является RUSMARC, который базируется на использовании международного коммуникативного формата UNIMARC с учетом трактовки и категорий действующих в России ГОСТов и правил каталогизации.

Современные АБИС позволяют экспортировать из них библиографические записи с использованием различных протоколов, осуществляя конвертацию в один из нескольких коммуникативных форматов «на лету». В систему интеграции библиографических записей, являющуюся частью Национальной электронной библиотеки, загрузка из АБИС происходит с использованием протокола OAI-PMH. При этом записи загружаются в формате MARCXML.

Задача публикации библиографических данных в открытое семантическое пространство

В 2006 году Тим Бернерс-Ли сформулировал четыре основных принципа связанных данных [2]:

1. Применение универсальных идентификаторов (Uniform Resource Identifiers - URI) в качестве имен сущностей.

2. Применение HTTP URI для 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 представления данных RDF (Resource Description Framework) [7], которая пригодна для машинной обработки. Структурно выражения в RDF представляют 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 [8].

Основная идея RDF состоит в том, чтобы показать взаимосвязь одних данных с другими. Объекты одного триплета RDF могут являться субъектами другого триплета, что позволяет рассматривать множество триплетов как ориентированный граф. Вершинами в таком графе являются субъекты и объекты, а ребрами - предикаты.

RDF представляет собой абстрактную модель представления данных с помощью триплетов и никоим образом не затрагивает семантики описываемых данных. RDF не предоставляет никаких терминов для описания классов вещей из реального мира и того, как они соотносятся друг с другом. Для выражения семантики используются словари, таксономии и онтологии, которые задаются с использованием языков RDFS (RDF Vocabulary Description Language, более известный как RDF Schema), SKOS

ШОРИН Олег Николаевич *- заместитель генерального директора по информатизации ФГБУ «Российская национальная библиотека». Адрес: 191069, г. Санкт-Петербург, ул. Садовая, 18 e-mail: shorin@nlr.ru*

(Simple Knowledge Organization System) и OWL (Web Ontology Language) [9], соответственно.

Возможность использования терминов из уже существующих словарей является одним из основополагающих механизмов связанных данных. В сообществе уже разработан ряд словарей, которые описывают широко распространенные понятия. По возможности надо стараться использовать уже существующие словари и онтологии. В случае же, если существующие словари не позволяют в полной мере описать предметную область, необходимо разработать новый словарь, следуя набору рекомендаций, основанных на опыте пользования связанными данными [1]. В связи с этим одной из первостепенных задач является разработка онтологии предметной области.

С использованием принципов, предложенных Тимом Бернерсом-Ли, в интернете реализуется проект открытых связанных данных (Linked Open Data - LOD), целью которого является интеграция данных, информации и знаний посредством глобальных идентификаторов ресурсов URI и модели данных RDF. LOD представляет собой гигантский глобальный граф, состоящий из миллиардов RDF-триплетов. Эти триплеты содержат информацию из различных областей человеческого знания, в том числе библиографические сведения.

Согласно принципам LOD RDF-триплеты должны быть связаны друг с другом. При этом приветствуется наличие как можно большего количества связей между данными различных организаций. Очевидно, что наличия связей только между данными, полученными из библиографических записей НЭБ, недостаточно. Необходимо организовать связывание с уже имеющимися библиографическими данными в LOD.

Преобразование данных

Для публикации библиографических записей в LOD их необходимо преобразовать в формат RDF в соответствии с web-онтологией. Как уже говорилось выше, при создании новых словарей и онтологий надо максимальным образом использовать уже существующие. Среди существующих онтологий наиболее подходящей, на первый взгляд, выглядит применение MODS/RDF, поскольку для преобразования библиографических записей из формата MARCXML в MODSRDF можно использовать последовательность XSLT-шаблонов, разработанных Библиотекой Конгресса США. Однако у такого очевидного подхода существует недостаток, который в долгосрочной перспективе может привести к значительной деградации производительности при работе с опубликованными в LOD данными.

Как известно, MODS/RDF основан на схеме MODS, которая имеет древовидную структуру. Из-за этой древовидности в RDF-графе, который получается из MODS, возникает множество пустых узлов. Наличие пустых узлов не противоречит синтаксисам RDF, RDFS, OWL, SPARQL. Однако наличие пустых узлов в LOD влечет за собой ряд практических трудностей [6].

В частности, пустые узлы в RDF-графе могут быть помечены разными способами в зависимости от программного обеспечения, используемого для работы с RDF-графом. Эти пометки могут также изменяться с течением времени. Таким образом, для проверки, являются ли два RDF-графа одинаковыми, необходимо будет решить задачу изоморфности двух графов [3], которая принадлежит к классу NP. Также наличие пустых узлов создает сложности при составлении SPARQL-запросов, т.к. отсутствует возможность задания переменных для пустых узлов с последующим использованием этих переменных во вложенных или последующих SPARQL-запросах.

Существуют способы борьбы с пустыми узлами, например, сколемизация - синтаксическая трансформация, при которой пустые узлы замещаются новыми именами таким образом, что два трансформированных графа будут эквивалентны тогда и только тогда, когда исходные RDF-графы изоморфны [5]. Однако алгоритмы сколемизации имеют экспоненциальный рост сложности, что делает их малопригодными для практического применения в LOD. В связи с этим Консорциум Всемирной паутины рекомендует воздерживаться от использования пустых узлов при публикации данных в LOD [4].

Модель данных, используемая в Europeana, Europeana Data Model, является избыточной, поскольку основные задачи, решаемые этой моделью, возникают из-за попытки представления различных видов объектов, в том числе и составных, имеющих несколько, возможно, противоречивых описаний. В случае НЭБ более логично использовать предыдущую версию модели данных, используемой в Europeana, Europeana's Semantic Elements (ESE), которая имеет обратную совместимость со словарем Dublin Core Metadata Initiative (DCMI) Metadata Terms.

При детальном изучении информации, содержащейся в библиографических записях НЭБ, выяснилось:

1. Расширения, добавленные в ESE по сравнению с Dublin Core, не будут задействованы для отображения информации о произведениях.

2. В ESE отсутствуют элементы для отображения информации об авторах. Необходимые элементы присутствуют в схеме EDM и представляют собой элементы словаря Friend of a Friend - FOAF.

Получилось, что Dublin Core и FOAF позволяют отобразить всю информацию, содержащуюся в библиографических записях. К тому же, использование схем Dublin Core и FOAF для отображения информации из библиографических записей НЭБ обеспе-

чивает совместимость с огромным количеством данных, представленных в LOD.

Таким образом, схема, отображающая информацию из библиографических данных НЭБ, использует следующие пространства имен:

- dc: http://purl.org/dc/elements/1.1/
- rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
- foaf: http://xmlns.com/foaf/0.1/.

Схема состоит из двух классов: «Автор» и «Произведение». Класс «Автор» имеет следующие свойства: foaf:name и rdf:type. А класс «Произведение» - dc:title, dc:description, dc:subject, dc:type, dc:format, dc:language, dc:publisher, dc:date, dc:identifier, dc:creator, dc:contributor.

Поскольку библиографические записи хранятся в формате MODS, который является XML-документом, логично представлять связанные данные в формате RDF/XML, так как преобразование из формата MODS в RDF/XML можно осуществить стандартными средствами XML с использованием XSLT-шаблона. Подобный XSLT-шаблон преобразования данных в соответствии с вышеописанной онтологией был разработан в рамках проекта интеграции библиографических записей в открытое семантическое пространство.

Связывание данных

Среди организаций, которые уже опубликовали свои библиографические данные в LOD, можно отметить две крупные библиотеки - Библиотеку Конгресса США и Британскую национальную библиотеку. Среди небиблиотечных организаций выделяется проект DBpedia, направленный на извлечение информации из данных, созданных в рамках проекта Wikipedia, поскольку DBpedia содержит большое количество связанных библиографических данных. Связывание данных НЭБ с данными этих трех организаций позволяет говорить об успешной публикации данных в LOD.

Поскольку в данных, полученных из библиографических записей НЭБ, выделяются две сущности - автор и произведение, - то в качестве первого шага для связывания данных в LOD находятся те же самые произведения и авторы и между ними устанавливается соответствующая связь. Впоследствии могут быть написаны роботы, которые, анализируя связи между различными данными в LOD, могут в автоматическом режиме создать новые связи.

В широко применямых в LOD онтологиях наиболее популярными связями для одних и тех же объектов являются owl:sameAs и rdfs:seeAlso. Связь owl:sameAs ставится между URI двух объектов и означает, что на самом деле это один и тот же объект. Связь rdfs:seeAlso является более слабой по сравнению с owl:sameAs. Если субъект находится в отношении rdfs:seeAlso с объектом, то это означает, что объект может обеспечить дополнительную информацию о субъекте. Из определения этих связей очевидно, что owl:sameAs является симметричной и транзитивной связью в отличие от rdfs:seeAlso.

Таблица 1

Синтаксисы запросов

Источник	Синтаксис запроса
Библиотека Конгресса США	http://www.loc.gov/search/?all=true&q=<строка>
Британская национальная библиотека	http://bnb.data.bl.uk/search?object=<строка>
DBpedia	http://lookup.dbpedia.org/api/search.asmx/KeywordSearch?QueryClass=&QueryString=<строка>

У Библиотеки Конгресса США нет SPARQL-точки доступа, поэтому для поиска авторов и произведений необходимо пользоваться поиском через web-интерфейс.

При использовании SPARQL-точки доступа в полнотекстовом режиме на сайте Британской национальной библиотеки возникли проблемы. Ответ на SPARQL-запрос выдавался через большой промежуток времени, а иногда ответ вовсе не поступал. В связи с этим для поиска данных в Британской национальной библиотеке также пришлось воспользоваться web-интерфейсом.

У DBpedia есть хороший поиск по ключевым словам с выводом результата, отсортированного по «популярности» объекта, - сколько других сущностей ссылается на этот объект, поэтому логичнее использовать такой поиск, чем SPARQL-точку доступа.

Для каждого из трех сайтов на языке Java были созданы модули, каждый из которых осуществлял поиск по заданному выражению и возвращал найденные результаты. Синтаксис запросов для каждого из источников приведен в **таблице 1**.

Поскольку большая часть данных из Библиотеки Конгресса США и Британской национальной библиотеки представлена на английском языке, а в НЭБ преобладают данные на русском языке, то для повышения вероятности успешного поиска имена авторов и названия произведений предварительно переводились на английский язык с помощью сервиса Google Translate.

Для осуществления перевода также был создан отдельный модуль, реализованный на языке Java. Данный модуль обращался к онлайн сервису с использованием URL следующего вида:

https://translate.google.ru/translate_a/single?client=t&sl=auto&tl=en&dt=t&q=<строка>.

Таким образом, для каждого свойства foaf:name класса «Автор» и для каждого свойства dc:title класса «Произведение» с помощью разработанных модулей

осуществлялся перевод на английский язык, а затем поиск в трех разных источниках. Для каждого источника в качестве результатов поиска модуль возвращал множество претендентов на установление связей owl:sameAs и rdfs:seeAlso. Для отсечения заведомо неподходящих кандидатов использовался метод биграмм.

Исходное поисковое словосочетание и полученный результат проходят предварительную обработку, в рамках которой из выражений удаляются излишние пробельные символы, знаки пунктуации, слова приводятся в нижний регистр. После этого сравниваемые строки разбиваются на множества биграмм - двухбуквенные подстроки. Расстояние между строками вычисляется с использованием меры Жаккара, которая представляет собой соотношение количества совпадающих элементов множеств биграмм к суммарному количеству элементов в этих множествах:

$$J = \frac{|A \cap B|}{|A \cup B|}.$$

Если полученное значение меры оказывалось выше порогового значения T_1=0,95, то в RDF/XML-файл добавлялся RDF-триплет со связью owl:sameAs между объектами. Если полученное значение меры оказывалось меньше порогового значения T_1 и выше порогового значения T_2=0,7, то в RDF/XML-файл добавлялся RDF-триплет со связью rdfs:seeAlso между объектами. Пороговые значения T_1 и T_2 были установлены экспериментальным путем.

Среди всего массива библиографических записей Национальной электронной библиотеки количество уникальных значений «Автор» и «Название» меньше, чем количество записей. Количество уникальных значений поля «Автор» составляет 19,6% от общего числа записей. Для поля «Название» данное значение составляет 68,6%. Для того чтобы не осуществлять повторные запросы к сервису Google Translate, сайтам Библиотеки Конгресса США, Британской национальной библиотеки и DBpedia, алгоритм поиска подобных записей сохраняет в базе данных значение исходного поля, результаты своей работы, а также время, когда данная проверка была сделана. Таким образом осуществляется кэширование результатов запросов к онлайн-сервисам.

Для каждого нового поля алгоритм сначала ищет значение этого поля в базе данных. В случае присутствия данного значения в базе данных берется уже готовый результат работы алгоритма. Если же значение отсутствует в базе данных, то осуществляется поиск с использованием вышеописанного алгоритма. Подобная оптимизация алгоритма позволяет уменьшить количество запросов к онлайн-сервисам на 55,6%.

Поскольку сайты Библиотеки Конгресса США, Британской национальной библиотеки и DBpedia постоянно развиваются и пополняются новыми элементами, то существует вероятность, что по прошествии некоторого времени результат работы алгоритма окажется иным. Для одного и того же набора данных было сделано 2 независимых запуска алгоритма поиска подобных записей с разницей в 5 месяцев. Количество RDF/XML-файлов, для которых была установлена связь в более позднем эксперименте, больше количества RDF/XML-файлов с установленными связями из более раннего эксперимента на 4,66%. Вследствие постоянного расширения количества записей, представленных в семантической паутине, из базы данных результатов работы алгоритма ежедневно удаляются записи старше двух месяцев.

Производительность системы поиска подобных записей и установления связи с ними напрямую зависит от скорости соединения сервера с интернетом. По данным сервиса проверки пропускной способности интернет-соединения Speedtest.net, скорость соединения центрального сервера НЭБ с интернетом составляет 86,92 Мбит/с для скачивания файлов и 86,97 Мбит/с для загрузки данных на удаленный сервер. При такой скорости соединения среднее время поиска подобных записей и установления связей с ними составляет 0,17 секунды для одной записи.

Поскольку общее количество записей с уникальными полями составляет 18,8 миллиона, алгоритм поиска подобных записей и установления семантической связи обработал бы все библиографические записи Национальной электронной библиотеки за 888 часов, что превышает 1 месяц. Для ускорения процесса поиска подобных записей и установления связи с ними вышеописанный алгоритм выполняется параллельно 30 потоками, каждый из кото-

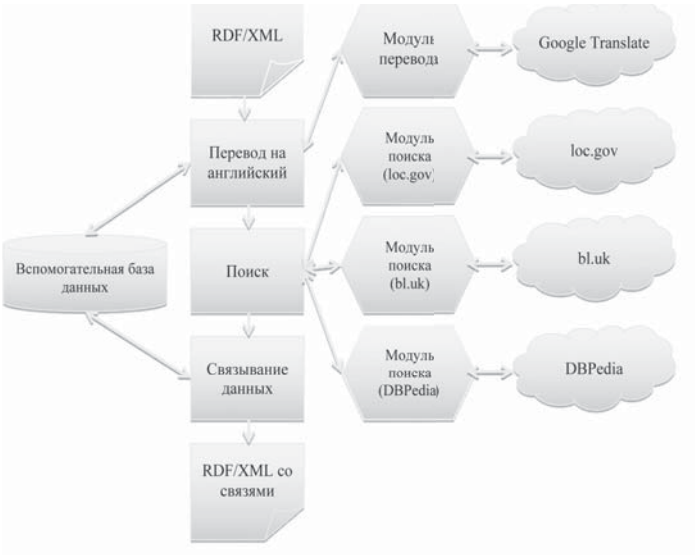

Рис. 1. *Архитектура системы семантического связывания данных*

рых обрабатывает независимое множество RDF/XML-файлов. Подобная оптимизация позволила сократить время работы алгоритма до 30 часов. Принципиальная схема работы системы поиска подобных записей и установления связей с ними приведена на **рис. 1.**

Заключение

С помощью разработанной системы связывания данных в автоматическом режиме было осуществлено связывание 33,8% RDF/XML-файлов. До начала работы алгоритма в каждом RDF/XML-файле содержалось в среднем по 4,46 RDF-триплета. После осуществления связывания данных среднее количество RDF-триплетов в RDF/XML-файлах с проставленными связями возросло до 15,16 RDF-триплетов. Распределение среднего количества RDF-триплетов связи в одном RDF/XML-файле в зависимости от ресурса и типа связи приведено в **таблице 2**.

Таблица 2

Среднее количество RDF-триплетов связи в одном RDF/XML-файле

	owl:sameAs	rdfs:seeAlso
Библиотека Конгресса США	4,04	2,51
Британская национальная библиотека	0,25	2,61
DBpedia	0,51	0,73

В качестве развития системы связывания данных можно выделить следующие направления:

1. Периодическое обновление связей для уже связанных RDF/XML-файлов, поскольку на сайтах Библиотеки Конгресса США, Британской национальной библиотеки и DBpedia появляются новые данные.

2. Создание полуавтоматической системы связывания данных, которая выдает варианты для установления связей, а решение о создании связи принимает человек.

3. Создание интеллектуальных систем, которые способны анализировать уже существующие связи между опубликованными в LOD данными и делать выводы о возможности связывания различных данных тем или иным видом связи.

Список литературы:

1. Allemang, D. Semantic Web for the Working Ontologist: effective modeling in RDFS and OWL / Dean Allemang, Jim Hendler. - Amsterdam: Elsevier; Morgan Kaufmann, 2008. - 352 p.: ill.

2. Berners-Lee, T. Linked Data. [Design issues] [Electronic resourse] / Tim Berners-Lee // W3C: World Wide Web Consortium. - Electronic data. - [Keio, Korea, Jap.; Cambridge, MA, USA; Biot, France; Beihang, China], 2006, last change: 2009. - Mode of access: https://www.w3.org/DesignIssues/LinkedData.html, free. - Title of screen.

3. Carroll, J.J. Matching rdf graphs / Jeremy J. Carroll // The Semantic Web - ISWC'2002: proc. first intern. Semantic Web conf., Sardinia, Italy, June 9 - 12, 2002 / ed. by: Ian Horrocks, James Hendler. - Berlin; Heidelberg: Springer-Verlag, 2002. - P. 5-15. - (Lecture notes in computer science; vol. 2342).

4. Cyganiak, R. RDF 1.1: concepts and abstract syntax: W3C Recommendation 25 Febr. 2014 [Electronic resourse] / Richard Cyganiak, David Wood, Markus Lanthaler // W3C: World Wide Web Consortium. - Electronic data. - Keio, Korea, Jap.; Cambridge, MA, USA; Biot, France; Beihang, China, 2004-2014. - Mode of access: https://www.w3.org/TR/rdf11-concepts/, free. - Title of screen.

5. Hogan, A. Skolemising blank nodes while preserving isomorphism [Electronic resourse] / Aidan Hogan // WWW'2015: proc. 24th intern. World Wide Web conf., May 18 - 22, 2015, Florence, Italy. - Electronic data. - [S. l.], 2015. - P. 430-440. - Mode of access: http://www.www2015.it/documents/proceedings/proceedings/p430.pdf, free. - Title of screen.

6. Hogan, A. Everything you always wanted to know about blank nodes / Aidan Hogan, Marcelo Arenas, Alejandro Mallea, Axel Polleres // J. of Web semantics: science, services and agents on the World Wide Web. - 2014. - Vol. 27/28. - P. 42-69.

7. Klyne, G. Resource Description Framework (RDF): concepts and abstract syntax: W3C Recommendation 10 Febr. 2004 [Electronic resource] / ed. by: Graham Klyne, Jeremy J. Carroll // W3C: World Wide Web Consortium. - Electronic data. - Keio, Korea, Jap.; Cambridge, MA, USA; Sophia Antipolis, France, 2004. - Mode of access: https://www.w3.org/TR/2004/REC-rdf-concepts-20040210/, free. - Title of screen.

8. Manola, F. RDF primer: W3C Recommendation 10 Febr. 2004 [Electronic resource] / ed. by: Frank Manola, Eric Miller, Brian McBride // W3C: World Wide Web Consortium. - Electronic data. - Keio, Korea, Jap.; Cambridge, MA, USA; Sophia Antipolis, France, 2004. - Mode of access: https://www.w3.org/TR/2004/REC-rdf-primer-20040210/, free. - Title of screen.

9. McGuinness, D. L. OWL: Web Ontology Language: overview: W3C Recommendation 10 Febr. 2004 [Electronic resource] / ed. by: Deborah L. McGuinness, Frank van Harmelen // W3C: World Wide Web Consortium. - Electronic data. - Keio, Korea, Jap.; Cambridge, MA, USA; Sophia Antipolis, France, 2004. - Mode of access: https://www.w3.org/TR/owl-features/, free. - Title of screen.