

Федеральное государственное бюджетное учреждение науки
Санкт-Петербургский институт информатики и автоматизации
Российской академии наук (СПИИРАН)

На правах рукописи

Шорин Олег Николаевич

**МЕТОДЫ И АЛГОРИТМЫ ИНТЕГРАЦИИ БОЛЬШОГО
ОБЪЕМА БИБЛИОГРАФИЧЕСКИХ ЗАПИСЕЙ В ОТКРЫТОЕ
СЕМАНТИЧЕСКОЕ ПРОСТРАНСТВО**

Специальность 05.13.11 – «Математическое и программное обеспечение
вычислительных машин, комплексов и компьютерных сетей»

Диссертация на соискание ученой степени
кандидата технических наук

Научный руководитель:
д. ф.-м. н., профессор
Серебряков В.А.

Санкт-Петербург
2016

Содержание

Введение	4
1 Анализ процессов интеграции библиографических записей в открытое семантическое пространство	13
1.1 Концепция Semantic Web	13
1.1.1 Web-онтологии	19
1.2 Обзор существующих решений	21
1.2.1 Europeana	22
1.2.2 Библиотека Конгресса	28
1.2.3 ЕНИП РАН	33
1.3 Постановка задачи и логическая схема исследования	40
2 Разработка методов и алгоритмов интеграции библиографических записей ..	43
2.1 Метод сбора библиографических записей	43
2.2 Алгоритм выявления дублетных библиографических записей и их слияние	47
3 Архитектура системы интеграции данных	70
3.1 Метод конвертации данных	70
3.2 Алгоритм связывания данных	78
3.3 Архитектура системы хранения данных и организации доступа	83
4 Апробация системы на электронном каталоге, консолидирующем библиографические записи из различных библиотек России	89
4.1 Национальная электронная библиотека	89
4.2 Результаты работы алгоритма выявления дублетных библиографических записей	93
4.3 Публикация библиографических записей НЭБ в Linked Open Data	95
Заключение	98
Список сокращений и условных обозначений	100
Список литературы	102

Приложения.....	123
Приложение А.....	123
Описание элементов словаря Europeana Semantic Elements.....	123
Приложение Б.....	141
Свойства основных классов Europeana Data Model	141
Уточнение свойств классов Europeana Data Model	144
Приложение В.....	145
Пример библиографической записи в формате MARC21	145
Пример библиографической записи в формате MARCXML	146
Пример библиографической записи в формате MODS	149
Пример библиографической записи в формате MODS/RDF	151
Свойства классов MODS/RDF.....	157
Приложение Г.....	163
Список библиографических записей, получаемых по протоколу OAI-PMH....	163
Приложение Д.....	170
XSLT-шаблон преобразования библиографических записей	170
Пример связанных данных в формате RDF/XML	182

Введение

Актуальность проблемы. Задача книгоописания возникла за несколько тысячелетий до нашей эры. В ту эпоху человечество начало создавать книжные собрания, которые требовали ведения некоего подобия каталога. Одним из старейших списков книг, который может быть признан каталогом, является шумерская глиняная плитка, датированная 2000 г. до н.э.. Она была найдена в Ниппуре и хранится сейчас в музее Пенсильванского университета [108]. Скорее всего, крупные книжные собрания появились в Древнем Египте гораздо раньше, но, к сожалению, папирус, используемый египтянами, не настолько долговечный материал, как глиняные таблички.

Мощный скачок в развитии каталогизация получила в середине VI века, когда в Европе стали появляться монастыри. Книги, хранившиеся в библиотеках монастырей, подлежали строжайшему учету, поскольку представляли собой значительную материальную ценность из-за дороговизны пергамента и огромной трудоемкости по переписке. С ростом фонда библиотек разрастался и каталог, что неизбежно привело к созданию элементарных правил, направленных на поддержание порядка в записях.

Постепенно в библиотеки начали поступать книги светского содержания, что повлекло за собой создание аппарата для описания книг, направленного на разыскание рукописей читателями. Описания в каталоге стали группироваться по различным признакам: именам авторов, форме произведений, предметам и характеристикам. В качестве примера такого каталога можно привести каталог Гладстонберийского аббатства [153].

Огромное значение на развитие книгоописания оказала попытка создания сводного каталога 183 монастырских библиотек Англии в 1250-1296 гг. Для обозначения библиотеки-участницы в каталоге использовалось краткое условное обозначение в виде номеров – сигл.

В XV в. значительно расширился круг авторов за счет произведений деятелей эпохи Возрождения. Изменился и состав читателей: помимо монахов

библиотекой стали пользоваться люди разносторонних интересов. Их запросы потребовали изменения правил каталогизации, что привело к появлению алфавитного авторского указателя к систематическому каталогу [131].

В середине 1440-х годов Иоганн Гуттенберг изобрел печатный станок [36], это привело к бурному развитию книготорговли. Каталоги стали выполнять функцию плакатов, рекламирующих издания на книготорговых ярмарках. В качестве примера можно привести афишу Петра Шеффера, которая была напечатана в 1469-1470 гг. и содержала 21 книгу.

Распространение книгопечатания поставило новые задачи перед каталогизацией: необходимо было охарактеризовать не отдельный экземпляр, а издание целиком. Для этого в описания книг были добавлены новые элементы – имена издателей, место и год издания.

Развитие науки и книгопечатания привели к бурному росту количества книг в библиотеках. В XVII в. стало очевидно, что библиотека представляет собой не просто набор книг. Она в обязательном порядке должна содержать четко организованную систему каталогов, позволяющую ориентироваться в фондах. Одним из первых такую мысль высказал Габриэль Ноде [115], который настаивал на необходимости создания алфавитного каталога, в то время как до этого времени считалось, что достаточно иметь только систематический каталог. Тогда же библиотеки начали издавать первые печатные библиотечные каталоги. В качестве такого примера можно привести каталог книг Бодлеянской библиотеки [44]. К XVIII в. библиотекари пришли к выводу, что печатать каталоги в виде книг неудобно, поскольку постоянный приток новых поступлений невозможно было оперативно отобразить в печатном каталоге. Появление карточного каталога позволило решить эту проблему – своевременно отображать в каталоге информацию о новых книгах.

Требования к библиотекам и их каталогам со стороны научного сообщества неуклонно возрастали, т.к. появлялись издания на иностранных языках, переводные издания, специфические фонды, например, географические карты, периодика, изобразительный материал. Каталоги библиотек не могли

удовлетворить требованиям читателей, что привлекло к проблемам каталогизации виднейших ученых XIX в. Именно в это время были заложены фундаментальные основы науки каталогизации в целом и формирования библиографической записи в частности.

Среди основных работ того времени хочется выделить исследования первого директора петербургской Публичной библиотеки А.Н. Оленина [37], Guillaume-François Debure [83], Antonio G. M. Panizzi [117], Thomas H. Horne [110], В.И. Собольщикова [45], Charles C. Jewett [103], Charles A. Cutter [80], Karl Dziatzko [85].

В первой половине XX в. развитие каталогизации было направлено в сторону создания национальных сводов правил книгоописания. А к середине прошлого столетия стало очевидно, что правила каталогизации в разных странах обязаны идти по пути неуклонного сближения [12]. На первый план вышли проблемы, связанные с использованием различных языков: транскрипция и транслитерация фамилий и имен, перевод заголовков и названий, описание переводных работ и международных договоров.

Во второй половине XX в. появились вычислительные машины, которые стали применяться в различных областях, в том числе и в каталогизации. Назрела необходимость создания машиночитаемого формата, который содержал бы перечень полей и правила их заполнения. Формальная логика вычислительных устройств требовала явного наличия или отсутствия элементов описания, в то время как человек мог интуитивно отделить автора от названия, глядя на каталогизационную карточку.

Разработкой и модернизацией форматов машиночитаемых библиографических записей в России занимались В.В. Скворцов [42], Е.И. Загорская [17, 24, 25], Н.П. Никольцева [17], Н.Н. Каспарова [5, 23–25], М.В. Экстрем, Т.А. Паршина, Г.Р. Ридер [59], Е.Д. Жабко [13], Т.Л. Масхулия [16, 35, 41], Ю.Г. Селиванова [16, 41], Л.В. Завьялова [16], М.В. Стегаева [16, 41], И.С. Дудник [5, 8], Т.А. Бахтурина [3–5], Г.П. Калинина [21, 22] и др.

Принципами создания автоматизированных библиотечных систем и электронных библиотек, опирающихся на использование машиночитаемых форматов библиографических записей, занимались Я.Л. Шрайберг [10, 11, 18, 57, 58], Ф.С. Воройский [9–11, 58], А.И. Земсков [18], Б.И. Маршак [33, 34], К.А. Колосов [28, 29], А.И. Племяк [38, 39], Б.Р. Логинов [32], Н.Е. Каленов, Г.И. Савин, А.Н. Сотников [20], В.М. Богданова, М.М. Горбунов-Посадов, Т.А. Полилова, М.И. Слепенков [7] и др.

В 2006 г. Tim Berners-Lee сформулировал основные принципы связанных данных [67], из которых на основе интернета можно построить семантическую паутину. Отличие семантической паутины заключается именно в термине «данные», которое ставилось в противовес существовавшему на тот момент, пусть и «гипер-», но всё же «тексту». С точки зрения связанных данных библиографические записи представляют огромный интерес, поскольку объекты, информация о которых хранится в записях, взаимосвязаны: 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.

Представлением библиографических записей в виде связанных данных занимаются ведущие каталогизаторы и специалисты информационных технологий В.В. Скворцов [43], О.Н. Жлобинская [14], А.А. Бездушный, А.Н. Бездушный, В.А. Серебряков, В.И. Филиппов [6], Bermès E. [66], Svensson L.G. [124, 126], Böhme C. [72], Tillet B. [129, 130] и др.

С использованием принципов, предложенных Berners-Lee, в интернете реализуется проект открытых связанных данных – Linked Open Data [71], целью которого является интеграция данных из различных областей знаний, в том числе и библиографической информации. Поставщиками данных из библиографических записей в LOD являются как отдельные библиотеки, так и различные консорциумы. В качестве примеров можно привести работы, ведущиеся в Библиотеке Конгресса, Национальной библиотеке Франции [123], Национальной библиотеке Германии [125], Британской библиотеке [136], проекте Виртуальный

Международный Авторитетный файл – Virtual International Authority File [95, 137], электронной библиотеке «Научное наследие России» [19].

Наличие открытого доступа к массиву библиографической информации с возможностью обнаружения семантически связанных данных является одной из составляющей развития как культуры в целом, так и отдельных направлений в книжной отрасли в частности. Преимущества от публикации библиографических записей в Linked Open Data трудно переоценить. На 78-ом Всемирном библиотечном конгрессе Международной федерации библиотечных ассоциаций IFLA, который состоялся в 2012 году в Хельсинки, были озвучены следующие преимущества от публикации данных в связанном виде [60]:

1. Открытый доступ и обмен метаданными.
2. Способствование случайному обнаружению новых источников данных.
3. Выявление основных шаблонов использования ресурсов и метаданных.
4. Навигация, основанная на использовании фасетов.
5. Обогащение метаданных с использованием ссылок.

Однако, при публикации данных, собранных из различных источников, неизбежно возникают вопросы интеграции данных: выявление дублетных записей и их слияние. Решением подобного рода проблем занимались Ivan P. Feleggi и Alan B. Sunter [86], Jeremy A. Hylton [88], Mauricio A. Hernández и Salvatore J. Stolfo [94], William E. Winkler [133], Pawel Jurczyk [105], В.А. Серебряков, А.Б. Антопольский, А.А. Каленкова, А.Н. Сотников [1], Д.Н. Рубцов и В.Б. Барахнин [40], О.М. Атаева и Л.Н. Шиолашвили [2], А.А. Князева, И.Ю. Турчановский, О.С. Колобов [27] и др.

В мире существует всего несколько проектов по решению комплексной проблемы интеграции данных из различных библиотек, представленных в разных форматах, с выявлением дублетных записей с последующим их слиянием, преобразованием в связанные данные и публикацией в LOD. Крупнейшим из них является проект по интеграции данных из европейских библиотек, музеев и архивов – Europeana [155]. Подход, применяемый в проекте Europeana, не может быть применен к решению нашей задачи, поскольку он основан [101] на

использовании единого формата представления данных и неавтоматизированном механизме связывания данных.

Цель диссертационной работы. Расширение возможностей интеграции библиографических записей в открытое семантическое пространство.

Решаемая научная задача. Разработка методов и алгоритмов интеграции больших объемов библиографических записей в открытое семантическое пространство.

Подзадачи. Реализация поставленной цели предполагает решение следующих подзадач:

1. Анализ существующих решений в области интеграции библиографических записей в открытое семантическое пространство.
2. Разработка совместимой с уже существующими онтологии предметной области с учетом полноты представленной в библиографических записях информации.
3. Разработка алгоритмов, обеспечивающих установление близости текстовых полей библиографических записей.
4. Разработка системы интеграции библиографических записей, позволяющей формировать, хранить и предоставлять доступ к данным с использованием принципов Linked Open Data.

На защиту выносятся:

1. Структура и формат онтологии для публикации данных, полученных из библиографических записей, в открытом семантическом пространстве.
2. Масштабируемый алгоритм установления близости текстовых полей в библиографических записях, основанный на методе разбиения текстовых значений на биграммы с последующим отсечением с использованием меры Жаккара.
3. Алгоритм обнаружения дублетных библиографических записей и аддитивного пополнения данных в LOD в автоматическом режиме с использованием разработанной онтологии предметной области.

4. Архитектура сбора, хранения и публикации библиографических данных, позволяющая в автоматическом режиме осуществлять сбор библиографических записей из различных библиотек России, конвертировать их в формат, пригодный для публикации в LOD, проводить аддитивное пополнение данных и устанавливать связи с уже опубликованными в LOD данными.

Научная новизна. Новизна первого результата состоит в том, что в отличие от существующих решений предложена онтология, состоящая из минимального количества классов и свойств, что позволяет использовать её в качестве базиса для построения более сложных словарей, сохраняя при этом совместимость с уже существующими решениями. Отличительной особенностью второго научного результата выступает набор оптимизаций, основанных на методе построения множеств биграмм из текстовых строк с последующим использованием полученных биграмм для подсчета меры Жаккара. Третий научный результат характеризуется тем, что предложена совокупность правил адаптации алгоритма в зависимости от количества и качества библиографических записей, что делает алгоритм масштабируемым. Оригинальность четвертого научного результата состоит в гибридном подходе использования централизованной и распределенной архитектуры, позволяющем масштабировать полученную систему на сколь угодно большой объем данных без потери качества получаемых результатов.

Методы исследований. В работе применялись методы сравнительного анализа, моделирования, классификации, непараметрической статистики, сопоставления строк, связывания записей, а также методы анализа и синтеза информационных систем.

Теоретическая значимость исследования состоит в развитии концепции представления библиографических записей из разнородных источников в виде связанных данных, а также в определении технологических принципов дальнейшего расширения списка поставщиков метаданных. Также в результате исследования были разработаны алгоритмы выявления дублетных

библиографических записей, создания обогащенной записи и связывания данных с уже опубликованными данными в LOD.

Практическая значимость и реализация результатов исследования заключаются в создании модульного программного комплекса, позволяющего консолидировать библиографические записи из различных библиотек, выявить дубликаты и произвести их слияние, сконвертировать метаданные, используя разработанную схему представления, и опубликовать их в связанном виде в LOD. Благодаря созданному программному комплексу международное библиотечное сообщество получило информацию о российских публикациях, которая обновляется в автоматическом режиме.

Цели и задачи исследования определили логику изложения материала и структуру диссертационной работы. Она состоит из введения, четырех глав и заключения. Каждая глава завершается выводами. Список литературы содержит 156 наименований. В диссертации 7 рисунков, 12 таблиц, 1 график, приводится 5 приложений.

Апробация работы. Основные положения диссертации изложены в 10 публикациях [47-56], две из которых [48, 52] опубликованы в изданиях, входящих в перечень ВАК РФ российских рецензируемых научных журналов, в которых должны быть опубликованы основные научные результаты на соискание ученых степеней доктора и кандидата наук по специальности 05.13.11. По теме диссертации были сделаны сообщения и доклады на международных научно-практических конференциях, симпозиумах и форумах: 79th IFLA General Conference and Assembly «IFLA World Library and Information Congress» (Сингапур, 2013г.), Двадцать первая Международная конференция «Крым-2014» «Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса» (г.Судак, Крым, Россия, 2014 г.), XIII Международная научно-практическая конференция «Электронный век культуры» (г.Сочи, Краснодарский край, Россия, 2014г.), Четвертый Всероссийский симпозиум «Инфраструктура научных информационных данных и систем» (г.Санкт-Петербург, Россия, 2014г.), 13-я Научно-практическая конференция «Участники и

пользователи Национального информационно-библиотечного центра ЛИБНЕТ» «ЛИБНЕТ-2014» (г.Звенигород, Россия, 2014г.), ЕМС Forum 2014 (г.Москва, Россия, 2014г.), 18-е заседание Совета сотрудничества национальных библиотек России (г.Санкт-Петербург, Россия, 2014г.), V Всероссийская научно-практическая конференция «Фонды библиотек в цифровую эпоху: традиционные и электронные ресурсы, комплектование, использование» (г.Санкт-Петербург, Россия, 2015г.), Международный профессиональный форум: «Книга. Культура. Образование. Инновации.» (г.Судак, Крым, Россия, 2015г.), XVII Всероссийская научная конференция «Научный сервис в сети Интернет» (пос.Дюрсо, Краснодарский край, Россия, 2015г.), Пятый Всероссийский симпозиум «Инфраструктура научных информационных данных и систем» (г.Санкт-Петербург, Россия, 2015г.), Второй Международный профессиональный форум: «Книга. Культура. Образование. Инновации.» (г.Судак, Крым, Россия, 2016г.).

Реализация и внедрение результатов работы. Разработанная в диссертации система семантического связывания библиографических связей внедрена и используется в ФГБУ «Российская государственная библиотека» - операторе Национальной электронной библиотеки (НЭБ), что подтверждено справкой о внедрении №77/11-1567 от 18.10.2016г. В процессе разработки алгоритм выявления дублетных библиографических записей с последующим их слиянием был также апробирован на массиве записей ФГБУ «Российская национальная библиотека».

1 Анализ процессов интеграции библиографических записей в открытое семантическое пространство

Настоящая глава посвящена обзору предпосылок возникновения семантической паутины, а также обзору существующих решений по семантической интеграции библиографических записей.

1.1 Концепция Semantic Web

Отцом современного интернета считается Тим Бернерс-Ли (Tim Berners-Lee), хотя, строго говоря, это не совсем так. Протоколы, с помощью которых компьютеры обменивались данными, были созданы в Агентстве по перспективным оборонным научно-исследовательским разработкам США (DARPA) в конце 60-х годов XX века.

В 1989 году Тим Бернерс-Ли предложил мировому сообществу концепцию Всемирной паутины. В то время он был сотрудником Европейского центра ядерных исследований (ЦЕРН), где он обратил внимание на разрозненность информации, хранившейся на разных компьютерах. Процесс передачи информации с одного компьютера на другой нередко сопровождался многочисленными сложностями, связанными как и с несовместимостями форматов, так и наличием иерархической структуры подчинения отделов.

Тим предложил использовать принцип перекрестных ссылок, чтобы исследователь мог получить информацию, находящуюся на другом компьютере, непосредственно с него, не копируя к себе и не проходя всё иерархическое дерево подчинения отделов друг другу. 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 последующих лет им были разработаны протокол передачи гипертекста HTTP, язык гипертекстовой разметки HTML, глобальные идентификаторы ресурсов URI. Предложенные им технологии используются человечеством и по сей день. Основная заслуга Тима Бернерс-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 «всемирной паутины» – World Wide Web.

Изначально интернет представлял собой среду, в которой общались технически образованные специалисты, ученые из различных областей науки: математики, физики, химии, биологи. Для создания и наполнения сайта требовались глубокие познания в программировании на различных языках, в том числе и HTML. На этом этапе созданием сайтов и их наполнением занималась небольшая группа людей. Контент создавался узким кругом высококвалифицированных специалистов и был предназначен для потребления профессионалами.

В конце XX века произошел знаменитый кризис пузыря доткомов, в результате которого было несколько подорвано доверие к интернету. Через несколько лет после этого возник термин Web 2.0. Основная суть Web 2.0 заключалась 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 в противовес устаревшему Web 1.0. Однако, большинство ученых, в том числе и Тим Бернерс-Ли, склонно считать, что Web 2.0 – это всего лишь удачный маркетинговый ход, а не новая технология, поскольку она базируется на тех же самых протоколах, которые использовались с самого начала существования интернета.

В любом случае, информация, представленная на сайтах, предназначалась для людей, поскольку основу интернета составлял гипертекст. Это означало, что основной смысл, значение скрывался в самом тексте и не существовало

формального способа извлечения этой сути, пригодного для автоматизированной обработки. 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. 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 [70].

Для оперирования данными необходимо было решить несколько ключевых вопросов [93]:

1. Каким образом обеспечить доступ к данным, чтобы их можно было повторно использовать?
2. Как должно происходить обнаружение данных, связанных с уже имеющимися данными?
3. Как приложения должны интегрировать разнородные данные, полученные из большого числа заранее неопределенных источников?

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

В 2006 году Тим Бернерс-Ли сформулировал четыре основных принципа связанных данных [67]:

1. Применение универсальных идентификаторов (Uniform Resource Identifiers - URI) [69] 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) [107], 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 [112].

Субъект в триплете указывается с помощью URI, который идентифицирует описываемый ресурс. Объект в триплете может иметь вид простого текстового значения: строка, число или дата; а может, как и субъект, указываться с помощью URI на другой ресурс, который имеет какое-либо отношение к субъекту. Предикат указывает какой тип отношений существует между субъектом и объектом. Предикат также идентифицируется с помощью URI, которые берутся из специальных справочников – коллекций URI.

В качестве примеров триплетов, полученных из одной библиографической записи, можно привести следующие:

Субъект	Предикат	Объект
Книга	Имеет название	Преступление и наказание
Книга	Написана автором	Достоевский, Ф.М.
Достоевский, Ф.М.	Годы жизни	1821-1881
Книга	Издана издательством	Мир книги Литература

Таблица 1: Триплеты RDF

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 рёбрами – предикаты (рис. 1).



Рисунок 1: Граф RDF

Поскольку URI, идентифицирующие субъекты, предикаты и объекты, уникальны, то все существующие связанные данные можно трактовать как один гигантский глобальный граф (Giant Global Graph - GGG) [68].

Использование RDF для связанных данных имеет ряд преимуществ:

1. Использование HTTP URI в качестве глобальных уникальных идентификаторов для субъектов, предикатов и объектов позволяет модели данных RDF использоваться в глобальном масштабе.
2. Каждый RDF-триплет является частью гигантского глобального графа, и каждый RDF-триплет может быть использован как стартовая точка для исследования связанных данных.

3. Модель данных RDF позволяет устанавливать связи между данными из разных источников.
4. Информация из разных источников может быть легко объединена с помощью слияния двух множеств триплетов в один граф.
5. RDF позволяет представлять информацию в одном графе, используя различные схемы, т.к. предикаты могут быть использованы из различных словарей.

Возможности модели данных RDF значительно шире, чем это необходимо для связанных данных. В связи с этим от некоторых возможностей RDF рекомендовано отказаться при представлении связанных данных. В частности, не рекомендуется:

1. Использовать RDF-коллекции, RDF-контейнеры и осуществлять RDF-овеществление, поскольку данные конструкции весьма громоздки.
2. Область использования пустых вершин должна быть ограничена документом, в котором они появились. Иначе значительно возрастает сложность слияния нескольких графов в один.

RDF не является форматом, а представляет собой лишь абстрактную модель для описания взаимоотношений между данными в виде триплетов. Для публикации RDF-триплетов в Web необходимо их сериализовать одним из нескольких существующих способов. Наиболее распространенным способом является представление в виде XML - RDF/XML. Синтаксис RDF/XML стандартизован [65] консорциумом W3C и широко используется для публикации связанных данных в интернете.

Для встраивания RDF-триплетов непосредственно в HTML-документы используют формат сериализации RDFa [132]. Изначально RDF информацию указывали в виде комментариев в HTML-документах, однако это оказалось не практичным, поскольку при изменении документа необходимо было четко контролировать согласованное изменение в комментариях. Впоследствии триплеты стали органично встраивать в объектную модель документа (Document Object Model, DOM). Формат сериализации RDFa широко используется в тех

случаях, когда владелец данных имеет возможность редактировать HTML шаблоны для публикации и не имеет права повлиять на технологию публикации данных.

Существует способ сериализации RDF, ориентированный на создание и чтение триплетов человеком – Turtle [64]. N-Triples является подмножеством Turtle, в котором отсутствует возможность использования пространства имен (namespaces) и других методов сокращения размера файла, например, компактные URI (CURIE) или вложенные конструкции. В связи с этим 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, поэтому файл N-Triples может быть считан и разобран построчно. К тому же избыточность файла N-Triples позволяет множеству алгоритмов архивации значительно уменьшать его размер. Эти два свойства сделали N-Triples основным стандартом при обмене большими объемами связанных данных, например, для создания архивных копий.

Множество современных языков программирования поддерживают нотацию JSON (JavaScript Object Notation), поэтому неудивительно, что существует способ сериализации RDF/JSON [82]. Публикация RDF-триплетов с использованием сериализации RDF/JSON позволяет множеству программистов, использующих такие языки как JavaScript, PHP, разбирать, анализировать и модифицировать связанные данные без установки дополнительного программного обеспечения и использования специализированных библиотек.

1.1.1 Web-онтологии

Как уже было сказано выше, 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) [74], SKOS (Simple Knowledge Organization System) [100] и OWL (Web Ontology Language) [113] соответственно.

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.

RDFS предназначен для описания простейших онтологий - словарей. Данные словари состоят из набора классов - `rdfs:Class` и их свойств – `rdf:Property`. Например, можно создать словарь, посвященный издательствам, и в нём описать класс «Издательство», который будет включать в себя все существующие издательства в виде членов этого класса. Также можно создать свойство «Расположено в» внутри класса «Издательство». С использованием этих терминов можно будет публиковать в виде связанных данных информацию об издательствах и о том, в каком городе они расположены.

RDFS также содержит механизмы для определения взаимоотношений между классами и свойствами:

- `rdfs:subClassOf` – указывает на то, что экземпляр одного класса также является экземпляром другого класса;
- `rdfs:subPropertyOf` – соответственно, указывает на то, что ресурс, обладающий одним свойством, также обладает и другим свойством;
- `rdfs:domain` – указывает на то, что любой ресурс, обладающий определенным свойством, является экземпляром одного или нескольких классов;
- `rdfs:range` – указывает на то, что все значения определенного свойства являются экземплярами одного или нескольких классов.

Использование вышеперечисленных механизмов позволяет получать логические следствия, которые не представлены в RDF-графе явно, но следуют из семантики. Например, если задать правило, что все издательства являются

юридическими лицами, то факт существования триплета «Мир книги Литература» является «Издательством» автоматически влечет за собой утверждение, что «Мир книги Литература» является «Юридическим лицом».

OWL расширяет возможности RDFS с помощью дополнительных механизмов. В частности, можно отметить:

- `owl:equivalentClass` и `owl:equivalentProperty`, которые предоставляют механизм для определения соответствий между терминами из разных словарей, что в свою очередь ведет к увеличению интероперабельности уже существующих словарей;
- `owl:inverseOf` – это свойство позволяет создателю словарю задекларировать, что одно свойство является инверсией другого. Например, если в графе имеется триплет «Достоевский, Ф.М.» является автором «Преступления и наказания», то можно логически вывести триплет о том, что «Преступление и наказание» написано «Достоевским, Ф.М.». Свойство `owl:inverseOf` является симметричным, поэтому в указанном примере из второго триплета можно получить первый.

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 [61].

1.2 Обзор существующих решений

С использованием принципов, предложенных Тимом Бернерсом-Ли, в интернете реализуется проект открытых связанных данных (Linked Open Data - LOD) [71], целью которого является интеграция данных, информации и знаний

посредством глобальных идентификаторов ресурсов URI и моделью 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.

Среди основных поставщиков библиографических данных в LOD можно выделить европейский проект интеграции информации о культурных ценностях библиотек, архивов и музеев Europeana, проект, осуществляемый в Америке Библиотекой Конгресса, а также российский проект Академии Наук по созданию Единого Научного Информационного Пространства. Далее в этой главе будут проанализированы словари, использующиеся в перечисленных проектах, на предмет возможного заимствования терминов.

1.2.1 Europeana

Europeana – это крупнейший в Европе проект по созданию цифровой библиотеки, который позволяет осуществлять поиск и получать информацию об объектах, представляющих культурную ценность, из различных архивов, библиотек, музеев, галерей. На конец 2014 года в Europeana было представлено более 36 миллионов объектов, поставщиками которых стали более 3000 организаций из 35 различных стран Европы [156].

В 2009 году внутри Europeana начались работы по обогащению метаданных и созданию связей в соответствии с принципами LOD как между элементами самой Europeana, так и с другими ресурсами LOD [88]. Поскольку Europeana собирает информацию из различных источников, то теоретически для одних и тех же объектов, например, имен или географических названий, имеются различные описания, которые дополняют и обогащают друг друга. Публикация этих данных в LOD делает возможным получение доступа к этим сведениям: люди и автоматические программы могут использовать эти данные, копировать, скачивать и распространять их.

Для того, чтобы данные из Europeana были опубликованы, поставщики обязаны сопровождать эти данные описаниями, которые соответствовали единой модели – Europeana’s Semantic Elements (ESE). За основу ESE был взят словарь Dublin Core Metadata Initiative (DCMI) Metadata Terms [84], который был дополнен 14 элементами, необходимыми для нужд Europeana. В Таблице №2 приведены элементы словаря ESE. Полное описание элементов ESE приведено в Приложении А.

Элемент	Уточняющий элемент
dc:title	dcterms:alternative
dc:creator	
dc:subject	
dc:description	dcterms:tableOfContents
dc:publisher	
dc:contributor	
dc:date	dcterms:created; dcterms:issued
dc:type	
dc:format	dcterms:extent; dcterms:medium
dc:identifier	
dc:source	
dc:language	
dc:relation	dcterms:isVersionOf; dcterms:hasVersion; dcterms:isReplacedBy; dcterms:replaces; dcterms:isRequiredBy; dcterms:requires; dcterms:isPartOf; dcterms:hasPart; dcterms:isReferencedBy; dcterms:references; dcterms:isFormatOf; dcterms:hasFormat; dcterms:conformsTo; europeana:isShownBy; europeana:isShownAt
dc:coverage	dcterms:spatial; dcterms:temporal

dc:rights	
dcterms:provenance	
europaana:country	
europaana:dataProvider	
europaana:language	
europaana:object	
europaana:provider	
europaana:rights	
europaana:type	
europaana:UGC	
europaana:unstored	
europaana:uri	
europaana:userTag	
europaana:year	

Таблица 2: Элементы словаря ESE

В 2011 году стало понятно, что для описания всего многообразия культурных ценностей, хранящихся в Europeana, словаря ESE недостаточно. Среди основных недостатков ESE были следующие:

1. Все метаданные приводились к единому подмножеству, мощность которого была невелика.
2. ESE не имела возможности описывать сложные по своей структуре данные, сводя всё к плоской модели.
3. Никким образом не учитывалась первоначальная природа данных, что приводило к потере ценнейшей информации.

Были начаты работы по созданию нового словаря – Europeana Data Model (EDM) [92]. Основные задачи, которые ставились перед EDM были:

1. Отделить объекты (картина, книга, представление и т.п.) от их цифровой копии.
2. Отделить объекты от их метаданных.

3. Позволить одному и тому же объекту иметь несколько описаний, которые потенциально могут иметь противоречивые сведения.
4. Добавить поддержку сложных объектов, которые состоят из других объектов.
5. Добавить поддержку контекстных ресурсов, используя концепцию контролируемых словарей.

Для преемственности было принято решение, что элементы ESE будут использованы в EDM, что позволит сохранить обратную совместимость с уже опубликованными данными. Для всех новых данных рекомендуется использовать модель EDM, но если поставщик ресурсов предоставит метаданные в модели ESE, то эти метаданные будут автоматически сконвертированы. Следуя правилам, принятым в LOD, при создании EDM максимально использовались уже существующие словари. В EDM используются следующие пространства имен:

- dc: <http://purl.org/dc/elements/1.1/>
- dcterms: <http://purl.org/dc/terms/>
- edm: <http://www.europeana.eu/schemas/edm/>
- ore: <http://www.openarchives.org/ore/terms/>
- owl: <http://www.w3.org/2002/07/owl#>
- rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
- foaf: <http://xmlns.com/foaf/0.1/>
- skos: <http://www.w3.org/2004/02/skos/core#>
- rdau: <http://www.rdaregistry.info/Elements/u/>
- wgs84_pos: http://www.w3.org/2003/01/geo/wgs84_pos#
- crm: <http://www.cidoc-crm.org/rdfs/cidoc-crm#>
- cc: <https://creativecommons.org/ns#>

Как уже было сказано, EDM разделяет понятия объекты культурного наследия и их цифровые копии. Этим EDM существенно отличается от ESE, хотя большинство элементов этих двух моделей совпадают. Для реализации разделения объектов и копий в EDM определены классы для описания базовых объектов – базовые классы. К основным базовым классам относятся:

- `edm:ProvidedCHO` – класс, описывающий объект культурного наследия;
- `edm:WebResource` – класс, описывающий веб-ресурс, который является цифровой копией;
- `ore:Aggregation` – агрегатор, который группирует классы вместе.

Например, книга «Преступление и наказание» может быть представлена классом `edm:ProvidedCHO`, а её отсканированная копия – `edm:WebResource`. Такое деление позволяет указывать соответствующие свойства для каждого класса. В частности, для класса `edm:ProvidedCHO` в качестве свойства `dc:creator` может быть указан «Достоевский, Ф.М.», а для `edm:WebResource` можно задать `dc:format` значением «pdf».

Поскольку метаданные об объекте культурного наследия и его цифровой копии разделены между двумя классами, должен быть механизм, позволяющий их как-то связать. Для этого и необходим класс `ore:Aggregation`. Он содержит свойства, позволяющие создать связи между различными классами, а также свойства, с помощью которых можно один раз описать общие между классами свойства для того, чтобы не описывать их в каждом классе по отдельности.

Для описания комплексных объектов в базовых классах определены дополнительные свойства, которые могут использоваться для отображения взаимосвязей между частями объектов. Например, `edm:ProvidedCHO` может иметь свойство `edm:isNextInSequence` для того, чтобы ссылаться на предшествующий объект. Используя такую возможность отображения взаимосвязей внутри объекта, провайдеры информации должны выбрать ту степень детализации описания, которая наилучшим образом подходит для их коллекции. Например, подшивка журналов состоит из отдельных журналов, а каждый журнал, в свою очередь, представляет собой сборник статей.

Вторая группа классов, определяемых в EDM, это контекстно-зависимые классы. В них хранится информация, которая может быть ассоциирована с объектами. В частности, к такой информации относятся данные из авторитетных файлов, контролируемых словарей, тезаурусов. Такие метаданные, вне всякого

сомнения, обогащают информацию об объектах, но их разумнее хранить отдельно. К основным контекстно-зависимым классам относятся:

- edm:Agent – хранит информацию о субъекте;
- edm:Place – хранит информацию о месте;
- edm:TimeSpan – хранит информацию о времени;
- skos:Concept – хранит информацию о действии;
- cc:License – хранит информацию о доступе и использовании.

В качестве примера использования контекстно-зависимых классов можно привести следующий: вместо указания свойства dc:creator «Достоевский, Ф.М.» для книги «Преступление и наказание» можно создать экземпляр класса edm:Agent, который будет предоставлять URI-ссылку на авторитетную запись в авторитетном файле, что даёт возможность использовать дополнительную информацию об авторе: перевод имени и фамилии на различные языки, информацию о датах жизни, ключевых событиях в биографии и т.п. Контекстно-зависимые классы – это основной путь обогащения данных в Europeana.

В Приложении Б приведены основные свойства вышеописанных классов модели данных EDM, а также иерархическая взаимосвязь между свойствами. С полным описанием классов и их свойств в EDM можно ознакомиться в [127].

По состоянию на июль 2015 года с использованием EDM в LOD опубликовано более 2.4 миллионов объектов, которые были получены от 8 поставщиков данных Europeana. Эти данные описывают объекты, хранящиеся в более чем 200 организациях, расположенных в 15 различных странах. Поскольку данные в LOD должны быть взаимосвязаны, то для публикуемых метаданных Europeana устанавливает ссылки на:

- Шведский портал культурного наследия – Swedish open cultural heritage (SOCH) [109];
- Информацию о географических местах – используются данные GeoNames [138];
- Метаданные об окружающей среде – используются данные General Multilingual Environmental Thesaurus (GEMET) [139];

- Информацию о людях – используются данные DBpedia [114];
- Периоды времени – используется словарь *adhos*.

1.2.2 Библиотека Конгресса

В 1961 году Библиотека Конгресса начала работы по созданию машиночитаемого формата каталогизации, и к 1966 году был разработан первый вариант формата MARC (Machine-Readable Cataloging). Количество MARC-форматов увеличивалось с каждым годом, и в 1990-е годы их число стало больше 50. В конце 1998 года, следуя тенденции унификации, Библиотека Конгресса США и Национальная библиотека Канады объявили о том, что форматы USMARC и CANMARC объединяются в один формат – MARC21. Дальнейшим развитием формата занимается сообщество, в состав которого входит множество организаций, однако Библиотека Конгресса по праву считается основным разработчиком этого формата.

В 2002 году специалистами Библиотеки Конгресса совместно с MARC Standards Office при участии независимых экспертов из библиотечного сообщества была разработана схема представления MARC21 записей в виде XML файла – MARCXML. Поля остались теми же самыми, что и в MARC21, поменялся только сам формат файла. Основным предназначением MARCXML являлся обмен данными между различными системами с использованием веб-сервисов.

Основные преимущества использования MARCXML:

1. Простая и гибкая схема представления MARCXML.
2. Конвертация из MARC21 в MARCXML осуществляется без потерь.
3. Имея MARCXML файл, можно однозначно восстановить MARC21 файл.
4. Для MARCXML файла достаточно создать таблицу стилей XML, чтобы отобразить только те элементы, которые представляют интерес.
5. Для редактирования записей можно использовать стандартные средства XML.

6. Конвертацию данных внутри полей можно осуществить с помощью стандартных механизмов XML.
7. Возможность осуществлять валидацию библиографической записи как по структуре, так и по содержимому.
8. Возможность создания новых утилит для работы с библиографическими записями сторонними разработчиками, без глубокого изучения специфики строения MARC21 формата.

В том же году в Библиотеке Конгресса была разработана еще одна схема описания метаданных для библиографических записей – Metadata Object Description Schema (MODS) [89]. MODS представляет собой XML схему, которая содержит в себе небольшое подмножество полей MARC формата и использует текстовые метки для обозначения полей, в то время как в MARC записи и MARCXML используются числа для обозначения полей и буквы для подполей. С одной стороны, подобное упрощение делает MODS более удобным для восприятия человеком. С другой, при обратной конвертации из MODS в MARC невозможно получить первоначальную запись, поскольку:

1. Часть данных утрачивается при конвертации из MARC в MODS, и их невозможно восстановить.
2. В некоторых случаях несколько разных полей MARC отображаются в одно и то же поле MODS. При обратной конвертации не представляется возможным восстановить информацию о том, в каком же именно поле содержалась информация.

Объектами в модели MODS выступают MODS ресурсы и MODS описания ресурсов. MODS ресурс – это любой ресурс, связанный с библиотекой, например, книги, журнальные статьи, фотографии, или изначально цифровые объекты. Это может быть физический или цифровой объект, или это может быть абстрактное понятие, такое как коллекция или работа.

Описание MODS ресурса включает описательные метаданные о MODS ресурсе, отношения между ним и другими MODS ресурсами и административные метаданные, связанные с описанием MODS ресурсов. MODS ресурс по

определению имеет хотя бы одно связанное MODS описание ресурса, ничто не мешает ресурсу иметь несколько описаний.

Когда MODS ресурс является физическим или цифровым объектом (в отличие от абстрактного понятия), описание MODS ресурса обычно включает в себя информацию, позволяющую осуществлять доступ к ресурсу. Информация в описании MODS ресурсов может поставляться по значению или по ссылке (включенной в описание или через ссылку на внешнее представление информации). Описание MODS ресурса можно рассматривать функционально, аналогично библиографической записи. MODS ресурс может быть связан с одним или несколькими другими MODS ресурсами. Например, MODS ресурс, который представляет собой коллекцию может иметь членов, которые сами являются MODS ресурсами. Как правило, эти отношения между MODS ресурсами будут выражаться как часть описания MODS ресурсов для одного или обоих соответствующих ресурсов.

Описательные метаданные для MODS ресурса агрегируют заглавия и имена, связанные с ресурсом наряду с субъектами и другими элементами данных, которые в дальнейшем помогут в описании ресурса. Для систем, описание MODS ресурсов и особенно его описательные метаданные помогают при индексации, поиске и отображения информации о ресурсах. Для пользователей, описание MODS ресурса помогает с идентификацией, поиском, выбором и доступом к MODS ресурсу. Административные метаданные предоставляют информацию об описательных метаданных, в первую очередь происхождения информации. Они включают в себя информацию, такую как персоны или организации, ответственные за описательные метаданные, и/или дата последнего изменения описательных метаданных, а также отношения, выраженные в описании MODS ресурсов.

Вместе с MODS Библиотека Конгресса развивает Metadata Authority Description Schema (MADS), которая представляет собой XML схему для описания метаданных об агентах (людях и организациях), мероприятиях,

географических местах, жанрах и т.д. MADS является дополнением к MODS для отображения авторитетных данных.

Для представления информации из MADS и MODS в виде связанных данных в Библиотеке Конгресса с 2010 года ведется работа по созданию онтологий MADS/RDF и MODS/RDF соответственно. Пространством имен MODS/RDF является <http://www.loc.gov/mods/modsrdf/v1#>. Также в MODS/RDF используются следующие пространства имен:

- madsrdf: <http://www.loc.gov/mads/rdf/v1#>
- owl: <http://www.w3.org/2002/07/owl#>
- rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
- rdfs: <http://www.w3.org/2000/01/rdf-schema#>
- xsd: <http://www.w3.org/2001/XMLSchema#>
- ri: <http://id.loc.gov/ontologies/RecordInfo>

Основные классы MODS/RDF приведены в таблице 3.

Область определения MODS элементов верхнего уровня	
Класс	Описание
modsrdf:ModsResource	Член этого класса является ресурсом MODS, описываемый описанием modsrdf. Этот класс является областью определения для свойств, соответствующих MODS элементам верхнего уровня.
Классы в пространстве имен MODS RDF, являющиеся областью значений свойств, областью определений которых являются ModsResource	
modsrdf:Cartographics	Класс агрегатор, являющийся областью значений свойств, соответствующих MODS картографическим элементам.
modsrdf:ClassificationGroup	Агрегирует классификационное число с его схемой. Предназначен для использования в тех случаях, когда схема не принадлежит

	контролируемому словарю.
modsrdf:IdentifierGroup	Агрегирует идентификатор с его типом. Предназначен для использования, когда тип не принадлежит контролируемому словарю.
modsrdf:Location	Класс агрегатор, являющийся областью значений свойств, соответствующих MODS элементам местоположения.
modsrdf: NoteGroup	Агрегирует заметку с ее типом. Используется только для заметок, для которых определен тип.
modsrdf:Part	Класс агрегатор, являющийся областью значений свойств, соответствующих MODS элементам частей.
modsrdf:RoleRelationship	Класс, который агрегирует имя с ролью.
Классы пространства имен MODS RDF , имеющие область определения на классах, перечисленных выше	
modsrdf:LocationCopy	Класс агрегатор, являющийся областью значений свойств, соответствующих MODS элементам копирования. Копия – это подэлемент of locationOfResource.
Классы пространства имен MADS RDF, имеющие область значений на MODS элементах	
madsrdf:ComplexSubject	Область значений для MODS субъекта с множественными категориями субъекта.
madsrdf:GenreForm	Область значений для MODS genre и MODS subjectGenre.
madsrdf:Geographic	Область значений для subjectGeographic, subjectGeographicCode, и MODS place.
madsrdf:Name	Область значений для name, subjectName, и publisher.

madsrdf:Occupation	Область значений для subjectOccupation.
madsrdf:Temporal	Область значений для subjectTemporal.
madsrdf:Title	Область значений для titlePrincipal, или titleUniform, когда нет основного имени, и MODS subjectTitle.
madsrdf:NameTitle	Область значений для MODS titleUniform, когда существует основное имя, так что может быть построено NameTitle.
madsrdf:Topic	Область значений для MODS subjectTopic.
Классы пространства имен RI для административных метаданных	
ri:AdminMetadata	Соответствует MODS recordInfo.

Таблица 3: Классы схемы MODS/RDF

В Приложении В приведены примеры библиографической записи в формате MARC21, MARCXML, MODS и MODS/RDF, а также свойства классов MODS/RDF.

С использованием технологии связанных данных Библиотека Конгресса опубликовала всю имеющуюся у нее информацию, которую можно было представить в виде словарей и авторитетных данных. Осуществляется активная работа по публикации других данных, хранящихся в Библиотеке Конгресса.

1.2.3 ЕНИИ РАН

Внутри Российской Академии Наук реализуется проект по созданию Единого Научного Информационного Пространства, основной целью которого является предоставление доступа к данным о научной деятельности сотрудников различных учреждений РАН, их научных достижениях, административной информации об организациях. Наличие такой системы позволит научным

сотрудникам получить информацию о смежных со своими работами, ведущихся в других коллективах.

Схема метаданных в ЕНИП состоит из 3 типов профилей:

- Базовый профиль, внутрь которого входит информация о предметных областях, применимых для всех учреждений РАН. На уровне этого профиля тематическая специализация отраслей науки является несущественной.
- Вспомогательный профиль, содержащий элементы, применимые не только к объектам научного мира, например, поддержка контролируемых словарей, тезаурусов, информация о географическом месторасположении, административно-территориальном делении, справочники национальных языков и т.п.
- Специализированные профили, которые содержат в себе элементы, используемые в специализированных научных сообществах.

В основной профиль ЕНИП входят четыре группы информационных сущностей:

1. «Участники научной деятельности». В эту группу входят сотрудники учреждений РАН, которые занимаются научной деятельностью. Сотрудники могут быть представлены как по отдельности в виде класса «Персона», так и в виде административных объединений с помощью классов «Организации» и «Подразделения», или же в виде неформальных объединений с использованием классов «Сообщества», «Рабочие группы» и «Коллективы».
2. «Научная деятельность». В эту группу входят классы «Научные мероприятия», которые подразделяются на «Симпозиумы», «Семинары», «Конференции». Также в эту группу входят «Проекты».
3. «Результаты научной деятельности». В этой группе представлены «Интернет-системы», «Базы данных», «Математические модели», «Изобретения» и т.п.

4. «Документы и публикации». В эту группу входят статьи, отчеты, научные труды и другие документы. В качестве примеров классов этой группы можно привести «Диссертации», «Распоряжения», «Монографии» и т.д.

Во вспомогательный профиль ЕНИП входят следующие группы элементов:

1. Элементы общего характера. В эту группу входят классы «Идентификатор», который используется для идентификации различных информационных ресурсов, и «Основной код УДК», выступающий в качестве элемента классификатора.
2. Элементы академического аспекта данных. Примерами классов этой группы являются «Рубрика ГРНТИ», «Рубрика РФФИ», «Специальность ВАК».
3. Элементы исторического аспекта данных, которые могут содержать историческую справку в произвольной форме.
4. Схема описания файлов. Основными классами являются «Файл данных», «Защищенный файл», «MIME-тип».
5. Схема контролируемых словарей и классификаторов. Эта группа представлена такими классами, как «Словарный термин», «Термин классификатора».
6. Схема тезаурусов. В группу входят классы «Термин тезауруса», «Аскриптор» и «Дескриптор».
7. Базовая и расширенная схемы описания территориального расположения объектов. В этих схемах представлены следующие классы: «Тип административно-территориальной единицы», «Геополитическая единица», «Административно-территориальная единица», «Страна».
8. Схема справочников национальных языков. Состоит из классов «Язык» и «Язык ISO-639».
9. Минимальная, базовая и расширенная схемы контактной информации. В эту группу входят такие классы как «Адрес», «Телефон», «E-mail», «Тип адреса» и «Тип телефона».

В специализированных профилях отображена следующая информация:

1. Библиографическая информация. Данный профиль отвечает за представление метаданных о публикациях – официально зарегистрированных печатных изданиях. Публикации подразделяются на несколько уровней:
 - а. Аналитический. Сюда входят публикации, которые являются частью какого-либо издания. Например, статья в журнале.
 - б. Монографический. Состоит из публикаций, представляющих собой один физический объект. Это может быть книга, выпуск журнала, том многотомника.
 - в. Сводный. На этом уровне осуществляется группировка публикаций из монографического уровня по какому-нибудь признаку. Например, многотомное издание или продолжающиеся периодические издания, такие как журналы. Группировка отдельных изданий на сводном уровне может быть осуществлена по коллекционному признаку: например, серия биографических книг «Жизнь замечательных людей».

Наличие различных уровней позволяет избежать дублирования информации, так как информация о журнале в целом и об отдельных выпусках заполняется один раз на сводном и монографическом уровнях соответственно. Если бы вместо иерархического подхода использовалась бы планарная модель, то при описании каждой статьи вместо проставления ссылки на выпуск журнала приходилось бы добавлять информацию и о конкретном номере журнала, и о журнале целиком. Это неминуемо привело бы к наличию ошибок.

2. Информация о разработках и продуктах. В некоторых случаях результатом научной деятельности является составной объект, который имеет ряд специфических для него атрибутов. Например, база данных имеет тематику и охват, программная система состоит из множества файлов, как исполняемых, так и исходных, каждый из которых может отличаться языком программирования, размером, требованиями к

настройкам компилятора. В профиль о разработках и продуктах входят метаданные, относящиеся к программному обеспечению, web-сайтам, базам данных, оборудованию и техническому обеспечению.

3. Математическая информация. Данный профиль обеспечивает отображение математических формул в любых текстовых атрибутах. Формулы могут просматриваться одним из нескольких способов: с использованием языка TeX [111], с помощью плагина MathML [62] или же в виде картинки. Также в профиле поддерживается возможность поиска по формулам, идентификация публикаций в различных системах реферирования, а также тематическая классификация ресурсов с использованием общепринятых классификаторов.
4. Конференции и поддержка проведения конференций. Конференции являются одним из общепризнанных форматов обмена информацией о научных достижениях. Проведение конференции представляет собой трудоемкий процесс, в который вовлечены различные лица и организации, каждый из которых отвечает за определенный аспект проведения конференции. В частности, должны быть организованы приглашения для участия в конференции, регистрация, сбор тезисов и докладов с последующим редактированием, составлена программа конференции, решены бытовые вопросы и т.п. Различные этапы подготовки и проведения конференции, а также роли её участников отображены в этом профиле.
5. Коллекции и архивы. Данный профиль предоставляет возможность описания коллекций – механизм группировки как физических, так и электронных документов. Данный профиль во многом опирается на подходы, используемые в UKOLN Research Support Libraries Programme (UKOLN RSLP CLD) [120], однако есть и своя специфика: в ЕНИП предлагается разделить этот профиль на базовую схему, которая используется в большинстве проектов, и на специализированные схемы, например, архивы.

6. Поддержка библиотечной деятельности. Данный профиль содержит информацию о единицах хранения, а также обо всех операциях, которые можно совершить с этими экземплярами. В частности, поддерживается возможность получения информации о месте хранения экземпляра, его наличии, возможности оставить заявку на получение экземпляра на руки с последующим уведомлением о готовности выполнения заявки. Также предусмотрены механизмы для заказа ксерокопий или сканирования отдельных страниц экземпляра, уведомления об истечении срока использования и возможности продления этого срока.

В ЕНИП используются следующие пространства имен:

Префикс	Пространство имен
Международные стандарты и предложения по наборам элементов метаданных:	
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns# – Словарь языка RDF
dc	http://purl.org/dc/elements/1.1/ – Словарь элементов Dublin Core Element Set v1.1
dcterms	http://purl.org/dc/terms/ – Словарь элементов Dublin Core Terms
dcmitype	http://purl.org/dc/dcmitype/ – Словарь элементов DCMI Types
dclib	http://dublincore.org/documents/library-application-profile/ – Словарь элементов на основе рабочей версии спецификации DC-Library Application Profile Working Draft [140]
vCard	http://www.w3.org/2001/vcard-rdf/3.0# – Предложения по представлению VCard в RDF [91]
prism	http://prismstandard.org/namespaces/1.2/basic/ – PRISM Element Set v1.2 [154]
pcv	http://prismstandard.org/namespaces/1.2/pcv/ – Схема контролируемых словарей PRISM

foaf	http://xmlns.com/foaf/0.1/ – Словарь элементов Friend of a Friend (FOAF) [75]
cld	http://purl.org/rslp/terms# – Словарь элементов описания коллекций UKOLN RSLP CLD
Пространства имен основных схем метаданных ЕНИП:	
core	http://umeta.ru/namespaces/blocks/foundation/ – Элементы основного сводного профиля
aux	http://umeta.ru/namespaces/blocks/auxiliary/ – Элементы общих вспомогательных схем
acd	http://umeta.ru/namespaces/blocks/aspects/academy/ – Элементы академического аспекта
hist	http://umeta.ru/namespaces/blocks/aspects/historic/ – Элементы исторического аспекта
Пространства имен специализированных схем метаданных ЕНИП:	
bib	http://umeta.ru/namespaces/blocks/specializations/bibliography/ – Элементы специализации описания библиографической информации
prod	http://umeta.ru/namespaces/blocks/foundation/products/ – Элементы специализации описания информации о разработках и продуктах
math	http://umeta.ru/namespaces/blocks/specializations/mathematics/ – Элементы специализации описания математической информации
conf	http://umeta.ru/namespaces/blocks/specializations/conference-management/ – Элементы специализированной схемы поддержки проведения конференций
lc	http://umeta.ru/namespaces/blocks/specializations/library-management/ – Элементы описания традиционной библиотеки и поддержки

	библиотечной деятельности
thesauri	http://umeta.ru/namespaces/blocks/auxiliary/controlled-vocabularies/thesauri/ – Элементы специализированной схемы описания тезаурусов профиля вспомогательных схем ЕНИП

Таблица 4: Пространства имен ЕНИП

Более подробно с устройством схемы ЕНИП РАН можно ознакомиться в [6]. Схема метаданных ЕНИП использовалась при создании электронной библиотеки «Научное наследие России» [19], основной целью которой является предоставление информации о биографиях выдающихся ученых России, внесших вклад в развитие науки, их исследованиях, публикациях.

1.3 Постановка задачи и логическая схема исследования

Развитие Linked Open Data происходит по нескольким направлениям, среди которых можно выделить два основных:

1. Разработка онтологий для различных предметных областей, совместимых с уже существующими онтологиями.
2. Публикация данных в LOD в соответствии с разработанными онтологиями.

В России накоплен огромный массив информации об опубликованных экземплярах произведений, который создан в процессе каталогизации в различных библиотеках. В то время, когда исследователи из разных стран активно развивают семантическую паутину в части библиографической информации, вклад России в процесс всемирной интеграции данных минимален. Банальное копирование применяемых другими странами решений для нашей страны не подходит в силу многих факторов: большого объема накопленной информации, наличия разных национальностей и использования множества языков и

диалектов, каждый из которых обладает своими особенностями, использования национального формата машинной каталогизации.

Такой многонациональной, многоконфессиональной, мультикультурной, огромной стране, как Россия, со своими специфическими особенностями, появившимися в силу исторических причин, для бесшовной интеграции в уже сложившуюся инфраструктуру требуется разработка методов и алгоритмов, позволяющих в автоматическом режиме произвести все необходимые преобразования данных.

Исходя из вышесказанного, можно сделать вывод, что в настоящий момент созрела необходимость решения следующих задач:

1. Разработать совместимую с уже существующими онтологию предметной области с учетом полноты представленной в библиографических записях библиотек России информации. При разработке онтологии особое внимание следует уделить тому факту, что библиографические записи создаются различными библиотеками, имеющими свои особенности каталогизации. Онтология должна быть спроектирована таким образом, чтобы количество пустых данных было минимальным вне зависимости от того, по каким правилам каталогизации была создана библиографическая запись.
2. Разработать алгоритмы, обеспечивающие установление близости текстовых полей библиографических записей. В силу географической распределенности России большая часть библиотек получает одни и те же произведения, что влечет к созданию похожих, но всё-таки отличающихся друг от друга полей в библиографических записях. Отличия обусловлены как различиями в принятых в отдельных библиотеках правилами каталогизации, так и человеческим фактором, например, опечатками.
3. Разработать систему интеграции библиографических записей, позволяющую формировать, хранить и предоставлять доступ к данным с использованием принципов Linked Open Data. Поскольку

объем накопленной информации исчисляется десятками миллионов записей, проведение каких-либо операций в неавтоматизированном режиме полностью исключено. Разработанная система должна быть модульной, поддерживать возможность наращивания дополнительным функционалом, масштабируемой и совместимой с уже используемыми автоматизированными системами.

Исходя из постановки задачи, логичным образом вытекает структурная схема исследования. В первую очередь должны быть разработаны онтология предметной области и алгоритмы, позволяющие устанавливать меру схожести текстовых полей для большого количества исходных данных. На базе разработанных алгоритмов должна быть создана программная система интеграции библиографических записей. На заключительном этапе исследования должна пройти апробация разработанной системы на электронном каталоге, содержащем библиографические записи из различных библиотек России.

Выводы главы 1

Как было показано, семантическая паутина – бурно развивающееся направление, над которым трудится большое количество исследователей во всём мире. Глобальная цель интеграции всех имеющихся данных в целом, и библиографических данных в частности, разбивается на несколько подзадач, каждая из которых должна решаться с учетом множества взаимосвязанных факторов. Изучив предпосылки появления открытого семантического пространства, а также проанализировав крупнейшие мировые проекты по интеграции библиографических данных, в данной главе была сформулирована задача, которая максимально учитывает накопленный опыт в этой области.

2 Разработка методов и алгоритмов интеграции библиографических записей

В настоящей главе приводится исследование различных методов интеграции библиографических записей из разнородных источников, а также приводится описание алгоритма выявления дублетных записей с использованием меры Жаккара с последующим созданием обогащенных записей.

2.1 Метод сбора библиографических записей

Обмен информацией между библиотеками – устоявшаяся практика. Подобный подход позволяет экономить время и количество усилий, затрачиваемых на каталогизацию, поскольку количество одних и тех же экземпляров книг в различных библиотеках велико. Существуют различные протоколы, с помощью которых можно обмениваться библиографической информацией. Самыми распространенными и поддерживаемыми практически всеми существующими АБИС являются протоколы Z39.50 и OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting).

Протокол Z39.50 был изначально создан в Библиотеке Конгресса США в начале 80-х годов XX века. Основным предназначением этого протокола являлась унификация поиска в различных АБИС, абстрагированная от структуры хранения информации, поисковых языков и правил составления запросов, выходных форматов [46]. До появления этого протокола доступ к АБИС осуществлялся с использованием протокола HTTP, который является базовым протоколом навигации по гипертекстовым документам, и никоим образом не оптимизирован для работы с библиографической информацией.

Протокол Z39.50 позволяет провести поиск библиографических записей в различных АБИС, используя один и тот же синтаксис. Этот протокол широко используется при создании метапоисковых систем – систем, позволяющих

произвести поиск одновременно в несколько разных АБИС и агрегировать полученные результаты. Использование этого протокола позволяет нивелировать различия в моделях хранения, запроса и получения записей в конкретных реализациях.

В конце 90-х годов разработчики совместно с библиотекарями сформулировали две основные проблемы, связанные с интероперабельностью цифровых хранилищ данных [63]:

1. Конечные пользователи должны были оперировать с различными поисковыми интерфейсами, предлагаемыми различными системами.
2. Отсутствовали механизмы совместного использования метаданных, которые были бы машинно-ориентированными.

Для решения этих проблем предлагалось 2 альтернативных пути развития:

1. Перекрестный поиск информации в различных хранилищах с использованием протокола Z39.50.
2. Аккумуляция данных на центральном сервере из разных источников данных.

Подход с использованием перекрестного поиска информации имеет один огромный недостаток: система существенно деградирует при увеличении количества источников, т.к. скорость работы всей системы равна скорости работы самого медленного звена. В экспериментах, проведенных Networked Computer Science Technical Reference Library (NCSTRL) было показано, что при увеличении количества источников более 100 система, построенная на перекрестном поиске, переставала быть работоспособной [63].

Было предложено создать новый протокол Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), который был бы предназначен для автоматического сбора метаданных из различных АБИС и аккумуляции их на центральном сервере. Протокол OAI-PMH не является конкурентом протоколу Z39.50, поскольку не предназначен для поиска. Основное его назначение – быстро, эффективно, с минимальными затратами собрать библиографические

записи вместе, которые впоследствии могут быть использованы для создания других сервисов.

Одним из основных требований протокола OAI-PMH является предоставление поставщиками XML метаданных в формате Dublin Core. Другое немаловажное требование – обязательная реализация запросов, основанных на дате последней модификации библиографических записей.

Поскольку структура создаваемой системы интеграции библиографических записей представляет из себя распределенную конфигурацию с выделенным центральным сервером, на котором агрегируются записи из различных источников, то наиболее подходящим протоколом для сбора информации является узкоспециализированный протокол, изначально предназначенный для решения именно этой задачи – OAI-PMH. В связи с вышеизложенным в разрабатываемой системе протокол OAI-PMH является основой метода сбора библиографических записей из АБИС различных библиотек.

Стоит отметить, что объединенный мировой каталог WorldCat на базе протокола OAI-PMH создал Java-сервлет – OAICat [135], который держатели метаданных могут адаптировать и установить на своём оборудовании для автоматического сбора обновленных записей центральным сервером WorldCat. Европейская цифровая библиотека Europeana также создала сервис REPOX [128], который основан на использовании протокола OAI-PMH в качестве базового протокола для сбора ресурсов с тысяч различных серверов, расположенных по всей Европе. Нельзя не сказать, что обновленная версия Сводного каталога библиотек России, известная под названием СКБР2, агрегирует записи с помощью протокола OAI-PMH [31]. Использование протокола OAI-PMH для аккумуляции библиографических записей из различных источников такими крупными мировыми проектами еще раз убеждает в правильности выбора протокола доступа.

Для осуществления сбора библиографических записей из различных библиотек были созданы модули интеграции для автоматизированных библиотечных интегрированных систем (АБИС), используемых в библиотеках:

Aleph, Ирбис, MarcSQL и OPAC-Global. Схема сбора библиографических записей приведена на рисунке 2.

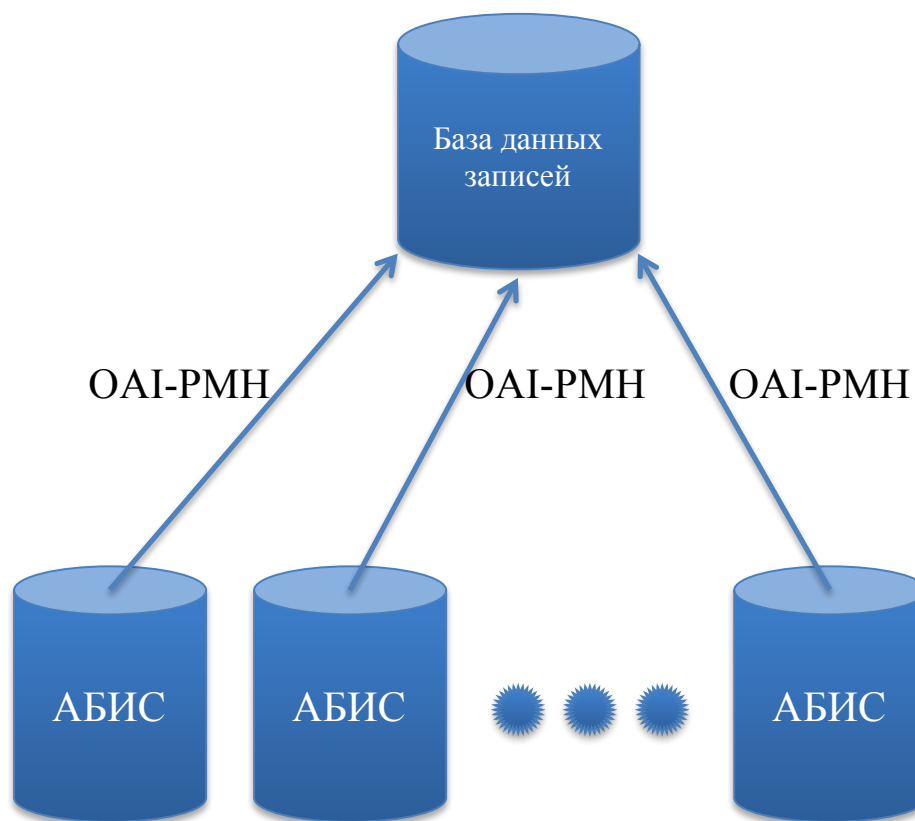


Рисунок 2: Сбор библиографических записей

Центральный сервер, на котором располагается хранилище данных, раз в сутки посылает запрос автоматизированным системам библиотек. Формат запроса выглядит следующим образом:

<http://aleph.nlr.ru/OAI?verb=ListRecords&metadataPrefix=marc21&from=2015-03-12&until=2015-03-12> . Параметрами from и until ограничивается временной период – одни сутки. В качестве ответа выдается список библиографических записей в формате MARCXML, измененных за указанную в запросе дату. Пример ответа, получаемого от ОAI-PMH сервера приведен в Приложении Г. Полученные от библиотек библиографические записи сохраняются в базе данных на центральном сервере для последующей обработки.

2.2 Алгоритм выявления дублетных библиографических записей и их слияние

Априорно известно, что при получении библиографических записей из разных библиотек, часто будет возникать ситуация, когда на одну и ту же сущность будет существовать несколько записей. Эти записи будут отличаться как по формату, так и по полноте заполнения, поскольку в различных учреждениях процессы каталогизации отличаются друг от друга. Например, состав дополнительных элементов, точек доступа может быть разным, системы классификации и предметизации, использование аббревиатур также могут разниться от одной организации к другой. К тому же, запись может содержать ошибки, опечатки, вызванные банальной человеческой оплошностью и невнимательностью.

Вопросы интеграции библиографических записей из различных источников с последующим их объединением и обогащением давно находятся в фокусе внимания ученых. Одними из основоположников этого направления являются Ivan P. Feleggi и Alan B. Sunter [86], которые разработали математическую модель связывания записей из двух различных множеств.

Предположим, что есть два множества A и B , состоящие из записей $a \in A$ и $b \in B$. Множество пар записей

$$A \times B = \{(a, b); a \in A, b \in B\}$$

разбивается на два подмножества:

$$M = \{(a, b); a = b, a \in A, b \in B\}$$

и

$$U = \{(a, b); a \neq b, a \in A, b \in B\}.$$

M – множество пар записей, описывающих один и тот же объект, а U – множество пар записей, описывающих разные объекты.

Сравнивая записи друг с другом, представляется возможным несколько вариантов вынесения решения о дублетности записей:

1. A_1 – положительное решение, т.е. $(a, b) \in M$;
2. A_3 – отрицательное решение, т.е. $(a, b) \in U$;
3. A_2 – возможное решение. В данных случаях невозможно однозначно сказать, к какому множеству M или U относится пара записей (a, b) .

Поскольку записи имеют сложную структуру, то для их сравнения используется не одно правило, а целый набор правил:

$$\gamma_i(a, b), \text{ где } i=1..K.$$

Каждое правило сравнения представляет собой функцию, принимающую в качестве параметров некоторые поля из записей a и b . Результатом работы функции является число, характеризующее степень схожести записей. Примером такого правила может служить следующая функция:

- при совпадении года издания в записях a и b , вернуть значение равное 1;
- при несовпадении года издания, вернуть 0;
- при отсутствии хотя бы в одной из записи информации о годе издания, вернуть 0,5.

Применив все правила, получим:

$$\gamma(a, b) = \{\gamma_1(a, b), \gamma_2(a, b), \dots, \gamma_K(a, b)\}.$$

Данное значение можно трактовать как точку в K -мерном пространстве.

Заранее выбрав пары (a, b) , для которых точно известно, что $(a, b) \in M$, можно построить множество точек

$$\Gamma^M = \{\gamma(a, b) | (a, b) \in M\}.$$

Это множество объединяет пары дублетных записей.

Аналогичным образом строится множество точек, объединяющих заведомо различные библиографические записи:

$$\Gamma^U = \{\gamma(a, b) | (a, b) \in U\}.$$

Таким образом, задача по выбору решения A_1 , A_3 или A_2 сводится к задаче вычисления расстояния от точки $\gamma(a, b)$ до множеств Γ^M и Γ^U .

Для подсчета расстояния в K -мерном пространстве используются различные функции, называемые метриками. В качестве примеров можно привести меру Евклида:

$$\sqrt{\sum_i (x_i - y_i)^2}$$

или меру Хемминга:

$$\sum_i |x_i - y_i|$$

В работе А.А. Князевой [26] для связывания библиографических записей в качестве функции $\gamma(a, b)$ используется комбинация точного сравнения и сравнения с усечением, определяемого с помощью стеммера Портера для русского языка [119]. А в качестве метрики для подсчета расстояния в K -мерном пространстве используется расстояние Махаланобиса.

Зачастую в библиографических записях содержатся ссылки на другие документы, например, полный текст произведения, которые бывают недоступны. В таких случаях приходится оперировать только теми строками, которые содержатся непосредственно в полях библиографических записей. В работе Д.Н. Рубцова и В.К. Барахнина [40] сравнение записей осуществляется только по полям «Автор» и «Название». В качестве функции $\gamma(a, b)$ используется вычисление длины наибольшей общей подпоследовательности двух строк (LCS – Longest Common Subsequence), нормированной к длине исходных строк. Таким образом, полностью совпадающие строки имеют значение 100%, а полностью различные – 0%. Авторы разработанной системы считают записи дублетными, если по полю «Автор» значение превышает 60%, а по полю «Название» - 80%.

Огромным достоинством авторов данной работы является предварительная обработка строк – преобразование букв с акцентами, а также перевод строк в нижний регистр. На заключительном этапе сравнения происходит проверка на исключения. В частности, разные части одной и той же статьи или книги могут иметь большую длину наибольшей общей подпоследовательности, но описывать при этом различные объекты. Для предотвращения таких случаев авторы

отдельно просматривают строки на наличие вхождений, обозначающих номера частей, например, «[1]», «Часть первая» и т.п.

Вышеописанная математическая модель имеет ряд недостатков, отмеченных самими разработчиками - Ivan P. Feleggi и Alan B. Sunter. Самым существенным из них является необходимость вычисления функции $\gamma(a, b)$ для всех пар записей $a \in A$, $b \in B$. В случае использования огромных массивов записей, полученных из разных источников, вычисление функции $\gamma(a, b)$ становится практически невыполнимым, поскольку сложность алгоритма равна $O(n^2)$.

Jeremy A. Hylton [98] предложил другой подход выявления дублетных записей. Основой его алгоритма является разделение записей на несколько кластеров. В кластер попадают записи, которые в терминах той или иной метрики располагаются недалеко друг от друга. Для выявления дублетных записей достаточно сравнить записи, входящие в состав одного кластера, что значительно снижает количество сравнений. Также в его работе показано, что процесс разбиения записей на кластеры должен предваряться процедурой нормализации – набором правил, применение которых приводит библиографические записи к некоему единообразному виду. К таким правилам нормализации можно отнести удаление избыточных пробельных символов, приведение строк к одному регистру, замена общепризнанных аббревиатур и правил написания на единообразный. Например, даты, записанные в разных форматах: Sept. 1987, 09.1987 – заменялись на какой-то один формат.

Алгоритм Jeremy A. Hylton состоит из двух этапов:

1. На первом этапе создаются кластеры записей, которые являются претендентами на дублетность.
2. На втором этапе записи из одного кластера сравниваются друг с другом для выявления дублетов.

Для разбиения всех записей на кластеры Jeremy A. Hylton использует полнотекстовый поиск. Для этого он выбирает 1 слово из поля «Автор» и 2 слова из поля «Название». Выбор слов происходит случайным образом. Далее алгоритм

осуществляет полнотекстовый поиск, используя выбранные слова, по тексту всех записей. Записи, попавшие в результат полнотекстового поиска, объединяются в кластер.

Все записи, попавшие в один кластер, сравниваются друг с другом, используя метод n -грамм. N -грамма – это последовательность из n символов. Например, для словосочетания «синтаксис и семантика» биграммами (n -грамма для n равного 2) с точки зрения слов являются «синтаксис и», «и семантика». С точки зрения символов все биграммы этой строки это: «си», «ин», «нт», «та», «ак», «кс», «си», «ис», «с », « и», «и », « с», «се», «ем», «ма», «ан», «нт», «ти», «ик», «ка».

Любое текстовое поле можно представить в виде вектора в пространстве размерности P^n , где P – количество символов алфавита, а n – размерность n -граммы. Элементами вектора является количество вхождений в строку той или иной n -граммы.

Например, для русского алфавита, дополненного символом пробела и арабскими цифрами, при разбиении на биграммы для каждого поля будет получаться вектор в пространстве размером 44^2

$$\vec{W} = \{s_{aa}, s_{ab}, s_{av}, \dots, s_{ш5}, \dots, s_{98}, s_{99}\},$$

где, например, $s_{пр}$ – количество вхождений биграммы «пр» в исходный текст.

Имея два вектора $\vec{W}_1$ и $\vec{W}_2$, полученными для полей из записей, претендующих на дублетность, можно посчитать расстояние между ними, используя одну из метрик. Если расстояние между векторами окажется ниже пороговой, то можно считать, что значения полей являются одинаковыми.

Jeremy A. Nylton в своей работе использует триграммы (n -граммы для n равного 3). В качестве алфавита он использует латинский алфавит, дополненный арабскими цифрами. Для вычисления расстояния используется мера Евклида. Таким образом, расстояние между векторами $\vec{A}$ и $\vec{B}$ вычисляется по формуле:

$$\|\vec{D}\| = \sqrt{\sum_{i=1}^{999} (a_i - b_i)^2}$$

Пороговое значение установлено экспериментальным путем с использованием двух выборок, состоящих из 50 записей, для которых априорно известно являются ли они дублетами или нет. Пороговое значение, применяемое Jeremy A. Hylton, равно

$$T = 2.486 + 0.025L,$$

где L – общее количество триграмм, полученное из двух строк.

Большим достижением Jeremy A. Hylton является не только снижение количества вычислений $\gamma(a, b)$ в алгоритме выявления дублетных записей, но и описание алгоритма создания обогащенной записи, содержащей объединенную информацию из нескольких записей, с последующим удалением исходных записей, содержащих менее полную информацию.

Слабым местом алгоритма Jeremy A. Hylton является первый этап: результаты полнотекстового поиска зависят от используемого алгоритма поиска, заранее построенного поискового индекса, а также от выбора ключевых слов, по которым осуществляется поиск.

Для применения алгоритма Jeremy A. Hylton к задаче выявления дублетных библиографических записей с их последующим слиянием необходимо существенно усовершенствовать первый этап алгоритма и адаптировать второй этап [47].

Перед началом работы алгоритма происходит нормализация библиографических записей:

1. Записи конвертируются из формата MARCXML в MODS с использованием XSLT-шаблона;
2. Последовательности пробельных символов заменяются на один пробел;
3. Все буквы приводятся к нижнему регистру;

4. Общеизвестные аббревиатуры и правила написания заменяются на единообразный формат.

Следующим шагом алгоритма является детерминированный поиск дублетов: если в разных записях совпадает ISBN (International Standard Book Number) – уникальный идентификатор книги, то эти записи описывают один и тот же объект. ISBN представляет собой последовательность цифр, разделенных дефисом или пробелом. Удалив дефисы и пробелы из ISBN, получается целое число, поиск по которым в сбалансированном дереве можно осуществить со сложностью $O(\log n)$.

После детерминированного поиска дублетов происходит вероятностный поиск, который использует подход Jeremy A. Hylton с существенными модификациями.

Перед разбиением всех записей на кластеры, они упорядочиваются по полю «Год издания». В данном случае используется следующее свойство: если у записей поле «Год издания» не является пустым, и записи являются дублетами, то у этих записей «Год издания» должен совпадать. Таким образом, отобрав претендентов на дублетность, нам достаточно будет сравнить только те записи, в которых «Год издания» либо совпадает, либо отсутствует. Данная оптимизация значительно снижает количество операций сравнения на втором этапе работы алгоритма.

Для разбиения всех записей на кластеры, в которые входят записи – претенденты на дублетность, используется функция отображения строки на множество чисел, при использовании которой значение функции не сильно изменяется при незначительном изменении аргумента. И наоборот, значение функции изменяется кардинально при значительном изменении аргумента.

Функции, отображающие текстовые строки на множество чисел, называются функциями хеширования и применяются в программировании достаточно давно. Однако, основными достоинствами широко распространенных функций хеширования является равномерное распределение по множеству чисел, поскольку зачастую данные функции используются в криптографии.

В нашем же случае используется функция хеширования, у которой незначительное изменение аргумента приводит к незначительному изменению результата функции хеширования. Подобные алгоритмы хеширования существуют, и они составляют семейство locality-sensitive hashing (LSH) функций.

Одно из определений семейства таких функций дано в работе P. Indyk и R. Motwani [99]. Согласно этому определению, семейство F функций $h: M \rightarrow S$ называется (R, cR, P_1, P_2) -чувствительным, если для порога $R > 0$, коэффициента $c > 1$, и любых двух точек $p, q \in M$ выполнены два условия:

1. Если расстояние $d(p, q) \leq R$, то $h(p) = h(q)$ с вероятностью не меньше P_1 ;
2. Если расстояние $d(p, q) \geq cR$, то $h(p) = h(q)$ с вероятностью не более P_2 .

Альтернативное определение семейства таких функций дано в работе M. Charikar [77]. Если sim – функции подобия объектов из множества P : $sim: P \times P \rightarrow [0, 1]$, то схема LSH – это семейство хеш-функций H , имеющих распределение D , таких что при выборе функции $h \in H$ согласно распределению D :

$$Pr_{h \in H}[h(p) = h(q)] = sim(p, q), \forall p, q \in P$$

Одним из наиболее часто используемых представителей из семейства LSH-функций является функция SimHash, поскольку ее реализация является простой. К тому же доказано [77], что мера Хемминга

$$\sum_i |x_i - y_i|$$

для двух хешей, посчитанных для двух множеств с использованием функции SimHash, пропорциональна коэффициенту Отиаи [116]

$$K(A, B) = \frac{|A \cap B|}{\sqrt{|A| * |B|}}$$

для этих множеств.

На практике для сравнения близости двух множеств чаще всего используется мера Жаккара [102]

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

В работе А. Shrivastava [122] доказано, что между мерой Жаккара J и коэффициентом Отиаи K для двух множеств выполнены следующие соотношения:

$$K^2 \leq J \leq \frac{K}{2 - K}$$

Таким образом, можно утверждать, что существует взаимосвязь между побитовым совпадением двух хешей, посчитанных для двух множеств с использованием функции SimHash, и мерой Жаккара для этих множеств. Именно это свойство функции SimHash и используется для разбиения библиографических записей на кластеры.

Шаги алгоритма, реализующего функцию SimHash, следующие:

1. Строка разбивается на биграммы, в результате чего для каждой строки получается множество биграмм.
2. Создаётся массив целых чисел, заполненный нулями, с размером, равным длине хеша в битах. В нашем случае длина хеша составляет 32 бита.
3. Для каждой биграммы из множества вычисляется хеш-функция. Обычно для строк используется функция хеширования, описанная Kernighan B.W. и Ritchie D.M. [73]:

$$\text{hash}(s) = s[0] * 31^{n-1} + s[1] * 31^{n-2} + \dots + s[n - 1].$$

Однако, в нашем случае вышеприведенная формула для вычисления хеша является малоприменимой, поскольку для биграмм она будет генерировать 16-битные хеши. Поэтому для генерации хешей используется широко известная функция MD5. Так как результат работы этой функции составляет 128 бит, в качестве хеша выбираются первые 32 бита.

4. Для каждого бита, полученного хеша, соответствующий ему элемент массива изменяется следующим образом: элемент массива

увеличивается на единицу в случае, если исходный бит равен 1 и уменьшается на единицу в противном случае.

5. На основе полученного массива генерируется результат хеширования – последовательность бит (в данном случае 32 бита, представленные типом `int`). Используется следующее соответствие элементов массива и хеша: если элемент массива больше нуля, то соответствующий бит равняется единице, иначе нулю.

Как можно заметить, данный алгоритм вычисления функции `SimHash` не зависит от порядка биграмм в исходном наборе данных. Используя вышеуказанный алгоритм, вычисляются хеши для множеств биграмм, полученных из строк полей «Автор» и «Название» всех библиографических записей. Таким образом, для каждой записи получается по два 32-битных хеша.

Поскольку количество записей велико, то сгенерировав по два хеша для всех записей, появляется проблема эффективного поиска среди этого набора тех хешей, расстояние между которыми по мере Хемминга мало, поскольку в худшем случае для поиска близких 32-битных хешей необходимо произвести 2^{32} операций побитового сравнения. А если хеши, полученные из строк полей «Автор» и «Название», рассматривать как один, то необходимо произвести 2^{64} операций сравнения.

Для снижения количества сравнений был использован метод разбиения хеша на равные части, описанный в работе Bingfeng Pi, Shunkai Fu, Weilei Wang и Song Nan [118]. В данной работе предлагается делить полученный хеш *hash* на равные части

$$hash_1 - hash_2 - \dots - hash_n.$$

Тогда поиск близких хешей можно осуществить с некоторой оптимизацией: среди имеющихся хешей находятся такие, в которых совпадает одна или несколько частей хеша.

В разработанном алгоритме способ разбиения хеша на части и метод сопоставления этих частей зависит от количества библиографических записей: чем больше записей, тем на большее количество кластеров должны быть

разделены все записи. Таким образом, в среднем в каждый кластер попадет небольшое количество записей. Это количество несоизмеримо мало по сравнению с общим количеством записей, поэтому при подсчете сложности работы алгоритма можно принять это число за константу, не зависящую от общего количества записей.

Для количества записей, измеряемых несколькими десятками миллионов, используется следующий алгоритм разбиения на кластеры: каждый из хешей полей «Автор» и «Название» разбивается на 4 части, и в качестве претендентов на дублетность выбираются записи, в хешах которых совпадает 2 из 4 частей для поля «Автор» и 2 из 4 частей для поля «Название». Т.е. для объединенного 64-битного хеша, имеющего вид

$$hash_1 - hash_2 - hash_3 - hash_4 - hash_5 - hash_6 - hash_7 - hash_8,$$

в один кластер попадут все записи, в которых совпадают части $hash_i - hash_j - hash_p - hash_q$, где $1 \leq i < j \leq 4$ и $5 \leq p < q \leq 8$.

Таким образом, используя вышеописанный алгоритм для поиска близких хешей для полей «Автор» и «Название», все исходное множество библиографических записей разбивается на кластеры, которые содержат претендентов на дублетность. Для всех записей каждого кластера используется метод сравнения строк, аналогичный алгоритму Jeremy A. Hylton. Однако, в отличие от его алгоритма разбиение происходит не с помощью триграмм, а с использованием биграмм – n -грамм, где n равно 2. Такой подход, во-первых, позволяет уменьшить размерность пространства, в котором получают векторы $\vec{W} = \{s_{aa}, s_{ab}, s_{av}, \dots, s_{ш5}, \dots, s_{98}, s_{99}\}$ для каждого поля. А во-вторых, для подсчета хешей использовались биграммы, поэтому и для подсчета меры должны использоваться биграммы.

Разбив поля библиографических записей на биграммы, можно посчитать расстояние между двумя записями, используя различные метрики. Как уже было сказано выше, в разработанной системе используется мера Жаккара, которая для двух множеств A и B представляет собой соотношение количества совпадающих

элементов этих множеств к суммарному количеству элементов в этих множествах:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Если трактовать множества биграмм для двух полей в виде векторов, то для двух векторов

$$\overrightarrow{W_1} = \{s_{aa}^1, s_{ab}^1, s_{av}^1, \dots, s_{ш5}^1, \dots, s_{98}^1, s_{99}^1\}$$

и

$$\overrightarrow{W_2} = \{s_{aa}^2, s_{ab}^2, s_{av}^2, \dots, s_{ш5}^2, \dots, s_{98}^2, s_{99}^2\}$$

мера Жаккара выражается формулой

$$J = \frac{\sum_{i=aa}^{99} (\min(s_i^1, s_i^2))}{\sum_{i=aa}^{99} (s_i^1 + s_i^2)}$$

Для записей, претендующих на совпадение, считаются меры Жаккара для множеств биграмм, полученных конкатенацией строк полей «Автор» и «Название». Если полученное значение выше порогового значения, а указанные записи не подпадают под исключения, описанные в работе Д.Н. Рубцова и В.К. Барахнина [40], то между этими записями устанавливается соответствие.

Теорема №1. Пусть $l(S)$ – длина строки S , а $Bi(S)$ – множество биграмм, полученных из строки S . Если для двух строк S_1 и S_2 выполнено неравенство

$$l(S_1) < 1 + k * (l(S_2) - 1)$$

то для меры Жаккара двух множеств, состоящих из биграмм этих строк, выполнено неравенство

$$J(Bi(S_1), Bi(S_2)) < k$$

Доказательство.

Заметим, что строка, длина которой составляет $l(S)$, разбивается на $l(S) - 1$ бигramму, то есть:

$$|Bi(S)| = l(S) - 1$$

Мера Жаккара двух множеств:

$$J(Bi(S_1), Bi(S_2)) = \frac{|Bi(S_1) \cap Bi(S_2)|}{|Bi(S_1) \cup Bi(S_2)|} = \frac{|Bi(S_1) \cap Bi(S_2)|}{|Bi(S_1)| + |Bi(S_2)| - |Bi(S_1) \cap Bi(S_2)|}$$

$$= \frac{|Bi(S_1) \cap Bi(S_2)|}{l(S_1) - 1 + l(S_2) - 1 - |Bi(S_1) \cap Bi(S_2)|}$$

Заметим, что если $|A| < |B|$, то $|A \cap B| \leq |A|$. Таким образом:

$$J(Bi(S_1), Bi(S_2)) \leq \frac{|Bi(S_1)|}{l(S_1) - 1 + l(S_2) - 1 - |Bi(S_1)|}$$

$$= \frac{l(S_1) - 1}{l(S_1) - 1 + l(S_2) - 1 - l(S_1) + 1} = \frac{l(S_1) - 1}{l(S_2) - 1}$$

Таким образом, неравенство

$$J(Bi(S_1), Bi(S_2)) < k$$

выполнено, если

$$\frac{l(S_1) - 1}{l(S_2) - 1} < k$$

$$l(S_1) - 1 < k * (l(S_2) - 1)$$

$$l(S_1) < 1 + k * (l(S_2) - 1)$$

Что и требовалось доказать. ■

Используя Теорему №1, реализуется дополнительная оптимизация, существенным образом сокращающая количество строк, для множеств биграмм которых необходимо считать меру Жаккара. Для записей, претендующих на совпадение, пары строк из полей «Автор» и «Название» проверяются на удовлетворение условию Теоремы №1 при k равном пороговому значению. Если обе пары удовлетворяют условию Теоремы №1 при k равном пороговому значению, то меры Жаккара для множеств биграмм, полученных из строк этих записей, не считаются, поскольку итоговая мера будет заведомо меньше порогового значения.

При подсчете меры Жаккара наиболее затратной операцией является вычисление мощности пересечения двух множеств, поскольку, как было показано выше:

$$J(Bi(S_1), Bi(S_2)) = \frac{|Bi(S_1) \cap Bi(S_2)|}{l(S_1) - 1 + l(S_2) - 1 - |Bi(S_1) \cap Bi(S_2)|}$$

Для оптимизации подсчета мощности пересечения двух множеств для каждой строки строится отсортированный список, состоящий из двух элементов:

1. Bigram - биграмма.
2. Count - количество раз, которое биграмма встречается в строке.

Такой список достаточно построить один раз для каждой строки, поскольку он может измениться только при изменении самой строки. Отсортировав подобный список по элементу с биграммами, алгоритм подсчета мощности пересечения множеств биграмм выполняется очень эффективно.

Теорема №2. Пусть $Bi(S)[j]$ – это j -тая биграмма в отсортированном списке биграмм, полученных из строки S . Для строки S_1 длиной $l(S_1)$ существует $r = [(1 - k) * (l(S_1) - 1)] + 1$ такое что, если $\forall p, q \leq r$ и строки S_2 выполнено условие $Bi(S_1)[p] \neq Bi(S_2)[q]$, то для меры Жаккара двух множеств, состоящих из биграмм этих строк, выполнено неравенство

$$J(Bi(S_1), Bi(S_2)) < k$$

Доказательство.

Заметим, что если для двух отсортированных списков биграмм $Bi(S_1)$ и $Bi(S_2)$ нет совпадающих элементов среди первых r биграмм, то мера Жаккара достигает своего максимального значения только в случае, если оставшиеся биграммы строки S_1 полностью совпадают с оставшимися биграммами строки S_2 . В таком случае мера Жаккара будет равна $\frac{|Bi(S_1)| - r}{|Bi(S_1)|}$. Потребовав, чтобы это отношение было меньше k , получим:

$$\frac{|Bi(S_1)| - r}{|Bi(S_1)|} < k$$

$$|Bi(S_1)| - r < k * |Bi(S_1)|$$

$$r > (1 - k) * |Bi(S_1)|$$

Как было замечено в Теореме №1, $|Bi(S_1)| = l(S_1) - 1$

Таким образом, ограничение выглядит следующим образом:

$$r > (1 - k) * (l(S_1) - 1)$$

Для уменьшения перебора нас интересует минимальное целочисленное значение r , которое равно $\lfloor (1 - k) * (l(S_1) - 1) \rfloor + 1$.

Что и требовалось доказать. ■

Таким образом, имея отсортированные списки биграмм для быстрого подсчета количества пересечений, можно использовать Теорему №2 для того, чтобы не проходить полностью списки для некоторых пар.

На языке программирования C схематический код алгоритма подсчета количества пересечений с использованием двух оптимизаций, описанных в теоремах, выглядит следующим образом (предположим, что указатель *head1* указывает на список, соответствующий строке S_1 , а указатель *head2* – на список, соответствующий строке S_2 , T – пороговое значение меры Жаккара):

```
result = 0;
optim1 = false;
optim2 = false;
r1 = 0;
r2 = 0;
l1 = strlen(s1);
l2 = strlen(s2);
if (l1 < l2) {
    r = (1 - T) * (l1 - 1) + 1;
    p = head1;
    q = head2;
    if (l1 < 1 + T * (l2 - 1))
        optim1 = true;
}
else {
    r = (1 - T) * (l2 - 1) + 1;
    p = head2;
    q = head1;
    if (l2 < 1 + T * (l1 - 1))
```

```

    optim1 = true;
}
if (!optim1) {
    while (p != NULL && q != NULL && !optim2) {
        if (strcmp(p->bigram, q->bigram) == 0) {
            result = result + min(p->count, q->count);
            p = p->next;
            q = q->next;
            r1 += p->count;
            r2 += q->count;
        }
        else {
            if (strcmp(p->bigram, q->bigram) < 0) {
                p = p->next;
                r1 += p->count;
            }
            else {
                q = q->next;
                r2 += q->count;
            }
        }
        if ((r1 > r) && (r2 > r) && (result == 0))
            optim2 = true;
    }
}

```

Результатом работы этого алгоритма может быть один из 3 вариантов:

1. Переменная *optim1* имеет значение «истина». Это означает, что сработала оптимизация, описанная в Теореме №1. Значение переменной *result* равно 0.

2. Переменная *optim2* имеет значение «истина». Это означает, что сработала оптимизация, описанная в Теореме №2. Значение переменной *result* равно 0.
3. И переменная *optim1*, и переменная *optim2* имеют значение «ложь». Это означает, что не сработала ни одна из оптимизаций. В переменной *result* содержится количество пересечений множеств биграмм.

В случае необходимости еще большего сокращения количества пар строк, для которых надо считать меру Жаккара, можно прибегнуть к методу построения обратных индексов, зависящих от длины строк, порогового значения меры Жаккара, позиции биграммы в строке, значения префикса и длины суффикса [134]. Построение подобного индекса потребует дополнительных затрат, но как и в случае создания отсортированного списка биграмм, этот процесс достаточно осуществить всего один раз.

Подсчитав мощность пересечений множеств биграмм и зная длину строк, вычисляется мера Жаккара. Как уже было сказано выше, если мера Жаккара превышает пороговое значение, то между этими библиографическими записями устанавливается соответствие. Пороговое значение было получено экспериментальным путем, и оно составляет:

$$T = 0,7$$

Для записей, между которыми установлено соответствие, запускается процесс создания обогащенной записи. Обогащение осуществляется по свойствам формата MODS. Для библиографической записи в формате MODS существуют простые и составные свойства. Для составных свойств происходит объединение наборов данных из различных записей. Для простых свойств возможны две ситуации. В случае, когда свойство присутствует в одной записи и отсутствует в других, берется свойство из той записи, в которой оно присутствует. В случае присутствия простого свойства в нескольких записях, используется подход, описанный Jeremy A. Hylton.

В соответствии с этим подходом, для каждого возможного написания свойства подсчитывается количество записей, в которых присутствует такое написание. В обогащенную запись попадает свойство, присутствующее в наибольшем количестве записей. Иногда различные написания присутствуют в одинаковом количестве записей. В таком случае в обогащенную запись попадает свойство, имеющее наибольшую длину, поскольку более длинное свойство содержит в себе больше информации.

Основываясь на опыте Консорциума музыкальных партитур библиотек Америки [79], который объединяет в себе информацию из 31 различной организации и содержит более 228 тысяч библиографических записей в формате MODS, было принято решение хранить на центральном сервере как обогащенные записи, так и первоначальные. Первоначальные библиографические записи могут пригодиться в случае поступления измененной, отредактированной записи из АБИС библиотеки. В этом случае для того, чтобы избежать коллизии, гораздо проще заново создать обогащенную запись, нежели выявить разницу между первоначальной и измененной записью и применить эту разницу к обогащенной записи.

Алгоритм выявления дублетных библиографических записей и создания обогащенных на их основе реализован на языке Java с использованием фреймворка Spring MVC, объектно-реляционного отображения Hibernate и библиотеки XMLBeans.

Выбор Spring MVC основан на том, что он позволяет создавать легко поддерживаемый исходный код за счет деления его на модули, соответствующие разным функциональностям. Подобное деление полностью согласуется с концепцией Model-View-Controller (MVC) [121].

Hibernate обеспечивает отображение классов Java на таблицы реляционных баз данных, таким образом освобождая разработчика от лишней работы. Этот подход не ограничивает разработчика исходно выбранной СУБД.

Для обработки данных в формате XML, используется библиотека XMLBeans. Она позволяет на основе схемы XML получать библиотеку или набор классов для представления данных, описанных в этой схеме.

Процесс выявления дублетов и слияния записей выполняется на центральном сервере системы. Данные берутся из хранилища, в которое попадают записи, собранные из различных библиотек с использованием протокола OAI-PMH.

В своей работе алгоритм использует временную базу данных, состоящую из 3 таблиц:

1. Таблица для хранения уникальных библиографических записей - BibRecords. Данная таблица состоит из следующих колонок:

Название колонки	Описание
id	Уникальный идентификатор библиографической записи
creation_time	Время создания записи
isbn	ISBN книги
xml	Содержимое библиографической записи
year	Год издания
distance	Значение меры Жаккара
author_string_id	Уникальный идентификатор хеша в таблице StringHash для поля «Автор»
title_string_id	Уникальный идентификатор хеша в таблице StringHash для поля «Название»
bigramm_list	Отсортированный список биграмм, полученных для поля «Автор» и «Название»
enriched_id	Уникальный идентификатор обогащенной библиографической записи в таблице EnrichedBibRecords

Таблица 5: Структура таблицы BibRecords

2. Таблица для хранения обогащенных библиографических записей - EnrichedBibRecords. Данная таблица состоит из следующих колонок:

Название колонки	Описание
id	Уникальный идентификатор
creation_time	Время создания записи
isbn	ISBN книги
xml	Содержимое обогащенной библиографической записи
year	Год издания
author_string_id	Уникальный идентификатор хеша в таблице StringHash для поля «Автор»
title_string_id	Уникальный идентификатор хеша в таблице StringHash для поля «Название»
bigramm_list	Отсортированный список биграмм, полученных для поля «Автор» и «Название»

Таблица 6: Структура таблицы EnrichedBibRecords

3. Таблица для хранения хешей - StringHash. Данная таблица состоит из следующих колонок:

Название колонки	Описание
id	Уникальный идентификатор хеша
value	Значение хеша
hash_part_1	Первые 8 бит хеша
hash_part_2	Вторые 8 бит хеша
hash_part_3	Третьи 8 бит хеша
hash_part_4	Последние 8 бит хеша

Таблица 7: Структура таблицы StringHash

В результате работы алгоритма обогащенные записи попадают в отдельную базу данных, также располагающуюся на центральном сервере. Общая схема

хранения исходных, вспомогательных и выходных данных на центральном сервере приведена на рисунке 3.

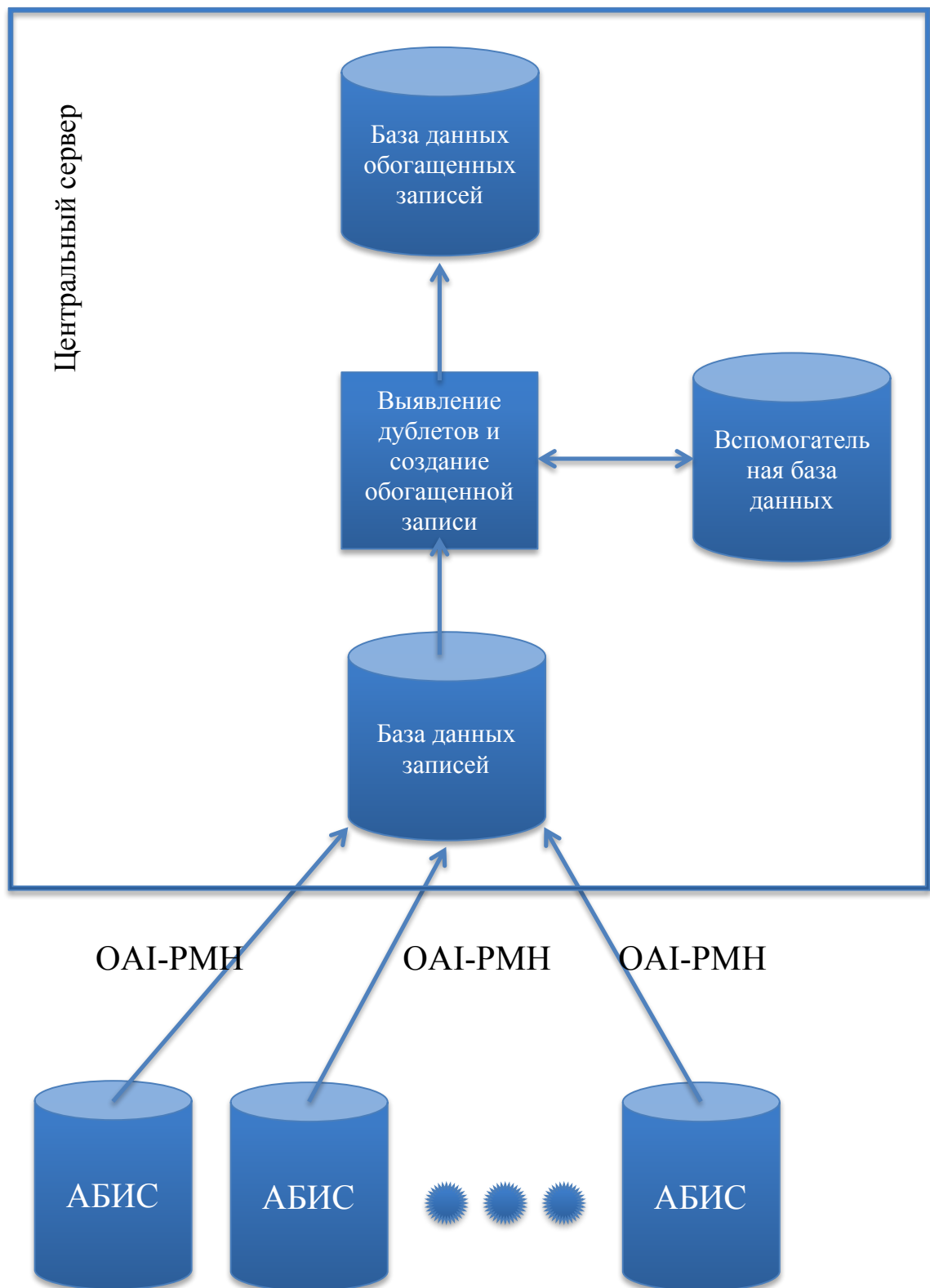


Рисунок 3: Схема хранения данных на центральном сервере

Для подсчета сложности алгоритма выявления дублетных библиографических записей следует рассмотреть три различных варианта: худший случай, лучший случай и средний случай. Обозначим через N – общее количество библиографических записей.

В худшем случае этап детерминированного поиска будет пропущен, поскольку ни у одной записи не будет заполнено поле, содержащее ISBN. Подсчет хешей для всех записей и построение отсортированных списков биграмм потребует $O(N)$ операций. В худшем случае все записи попадут в один кластер, таким образом потребуется вычислить меру Жаккара для всех возможных пар записей. Использование Теорем №1 и №2 может дать некоторую оптимизацию, но несмотря на это операция подсчета меры Жаккара для всех пар имеет сложность $O(N^2)$. Итого, общая сложность в худшем случае составляет $O(N^2)$.

В лучшем случае все библиографические записи будут содержать поле с ISBN, что позволит выявить все дубликаты на этапе детерминированного поиска. Сложность алгоритма в лучшем случае равна $O(\log N)$.

В среднем на этапе детерминированного поиска будут отобраны все книги и брошюры, имеющие ISBN. Согласно отраслевому докладу о книжном рынке России [30] в 2014 году было выпущено 112 126 названий книг и брошюр, среди которых много переизданий, переводных изданий и изданий, выходящих в сериях. Практика показывает, что в силу ряда причин в среднем ISBN содержат около трети библиографических записей электронного каталога. Таким образом, сложность детерминированного поиска составит $O(\log(0,33 * N))$, что эквивалентно $O(\log(N))$, а на этап вероятностного поиска попадет $0,67 * N$ записей.

Для каждой записи на этапе вероятностного поиска будет подсчитан хеш и построен отсортированный список биграмм – $O(0,67 * N)$ операций, т.е. $O(N)$. Как уже говорилось выше, разбиение на кластеры устроено таким образом, чтобы для каждой записи в среднем необходимо было выполнить фиксированное количество подсчетов меры Жаккара. Сложность данной операции составляет

$O(C * N)$. Таким образом, общая сложность алгоритма в среднем случае составляет: $O(N)$.

Выводы главы 2

Анализ широкораспространенных протоколов для интеграции библиографических записей показал, что для построения масштабируемых систем необходимо использовать протокол OAI-PMH. Для выявления дублетных библиографических записей в данной главе представлен алгоритм, основанный на вычислении меры Жаккара близости текстовых полей. Научной новизной предложенного алгоритма является ряд теорем, позволяющих оптимизировать самую вычислительно сложную часть алгоритма – подсчет количества пересечений двух множеств биграмм, полученных из текстовых строк. С практической точки зрения алгоритм является уникальным по сравнению с существующими за счет своей масштабируемости, которая достигается благодаря линейной сложности представленного.

3 Архитектура системы интеграции данных

В настоящей главе приводится описание выбранной онтологии для представления библиографических записей в LOD, подходы, применяемые для хранения связанных данных, а также алгоритм выявления уже опубликованных в LOD данных, пригодных для связывания.

3.1 Метод конвертации данных

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 могут 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 [144], который базируется на использовании международного коммуникативного формата UNIMARC с учетом трактовки и категорий действующих в России ГОСТов и правил каталогизации.

Библиографическая запись в формате RUSMARC имеет структуру, согласованную с международным стандартом ISO 2709. Каждая библиографическая запись должна содержать:

1. Маркер записи, состоящий из 24 символов;
2. Справочник;
3. Поля данных переменной длины, отделенные друг от друга разделителем поля.

С точки зрения формата поля данных могут быть обязательными и необязательными. К обязательным полям относятся:

- Идентификатор библиографической записи;
- Данные общей обработки;
- Заглавие и сведение об ответственности;
- Источник записи.

В зависимости от типа библиографической записи обязательными могут быть признаны и другие поля. Например, для документов, содержащих текстовую информацию, обязательным для заполнения является поле «Язык документа». К тому же, каждая конкретная библиотека может установить свои внутренние правила, согласно которым библиографическая запись будет считаться неполной без заполнения определенного набора полей, являющихся необязательными с точки зрения формата RUSMARC.

Количество библиотек, использующих RUSMARC в качестве коммуникативного формата, велико. Стоит отметить, что Сводный каталог библиотек России, позволяющий осуществлять корпоративную каталогизацию, использует формат RUSMARC.

Однако, помимо формата RUSMARC в России используются и другие разновидности формата семейства MARC. Например, Российская государственная библиотека использует формат MARC21, разработчиком которого является Библиотека конгресса США.

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 загрузка из АБИС происходит с использованием протокола OAI-PMH. При этом записи загружаются в формате MARCXML. Выбор этого формата был обусловлен следующими причинами:

1. АБИС Aleph, используемая в РГБ, не позволяет экспортировать записи в формате RUSMARC. В то время как, АБИС в других библиотеках способны выгрузить библиографические записи в формате MARCXML.
2. RUSMARC является национальным форматом и мало распространен за пределами России. Вследствие этого практически отсутствуют программные библиотеки и утилиты для работы с записями в этом формате. В противовес этому формат MARCXML представляет собой XML-файл, для которого применимы операции преобразования с использованием стандартных XML-инструментов. Существует достаточное количество шаблонов, инструментов, программных библиотек и утилит, активно поддерживаемых и развиваемых библиотечным сообществом, для работы с библиографическими записями в формате MARCXML.

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, описанных в первой главе, наиболее подходящей, на первый взгляд, выглядит использование 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 [97].

В частности, пустые узлы в RDF-графе могут быть помечены разными способами в зависимости от программного обеспечения, используемого для работы с RDF-графом. Также эти пометки могут изменяться с течением времени. Таким образом, для проверки являются ли два RDF-графа одинаковыми необходимо будет решить задачу изоморфности двух графов [76], 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 были изоморфны [96]. 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 [81].

Таким образом, MODS/RDF нежелательно использовать для представления библиографических записей, хранящихся в электронных каталогах библиотек России.

Схема метаданных ЕНИП РАН является избыточной в нашем случае, поскольку она направлена на отображение информации о научной деятельности сотрудников различных учреждений РАН. В связи с этим отображение библиографической информации в схеме ЕНИП РАН является подмножеством специализированного профиля, отвечающего за поддержку библиотечной деятельности. Использование избыточной схемы для отображения лишь малого подмножества информации из этой схемы является нежелательным.

Модель данных, используемая в Europeana, Europeana Data Model также 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 нашем случае 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 различных библиотек России, 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.

При более детальном изучении библиографических записей выяснилось, 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 российских библиотек обеспечивает совместимость с огромным количеством данных, представленных в LOD.

Таким образом, схема, отображающая информацию из библиографических данных, 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 2 классов: «Автор» и «Произведение». Класс «Автор» имеет свойства, представленные в таблице 8.

name	
Пространство имен	foaf
URI	http://xmlns.com/foaf/0.1/name
Название	Имя
Описание	Имя автора или название организации, являющейся автором произведения.
type	
Пространство имен	rdf
URI	http://www.w3.org/1999/02/22-rdf-syntax-ns#type
Название	Тип
Описание	Тип объекта: человек или организация.

Таблица 8: Свойства класса «Автор»

Свойства класса «Произведение» приведены в таблице 9.

title	
Пространство имен	dc
URI	http://purl.org/dc/elements/1.1/title
Название	Заголовок
Описание	Название произведения.

description	
Пространство имен	dc
URI	http://purl.org/dc/elements/1.1/description
Название	Описание
Описание	Текстовое описание произведения.
subject	
Пространство имен	dc
URI	http://purl.org/dc/elements/1.1/subject
Название	Предмет
Описание	Тема произведения.
type	
Пространство имен	dc
URI	http://purl.org/dc/elements/1.1/type
Название	Тип
Описание	Тип произведения: текст, издание, карты.
format	
Пространство имен	dc
URI	http://purl.org/dc/elements/1.1/format
Название	Формат
Описание	Формат файла, физический носитель или размер произведения.
language	
Пространство имен	dc
URI	http://purl.org/dc/elements/1.1/language
Название	Язык
Описание	Язык произведения.
publisher	
Пространство имен	dc
URI	http://purl.org/dc/elements/1.1/publisher

Название	Издатель
Описание	Персона, организация или сервис, ответственный за выпуск произведения.
date	
Пространство имен	dc
URI	http://purl.org/dc/elements/1.1/date
Название	Дата
Описание	Дата или период времени, ассоциированный со временем издания произведения, создания или оцифровки.
identifier	
Пространство имен	dc
URI	http://purl.org/dc/elements/1.1/identifier
Название	Идентификатор
Описание	Однозначный указатель на произведение: ISBN, ISSN и т.п.
creator	
Пространство имен	dc
URI	http://purl.org/dc/elements/1.1/creator
Название	Автор
Описание	Автор произведения. Представляет собой ссылку на класс «Автор».
contributor	
Пространство имен	dc
URI	http://purl.org/dc/elements/1.1/contributor
Название	Поставщик
Описание	Персона, организация или сервис, отвечающий за предоставление информации о произведении. Представляет собой ссылку на класс «Автор».

Таблица 9: Свойства класса «Произведение»

Библиографические записи хранятся в формате MODS, который является XML-документом. В связи с этим логично представлять связанные данные в формате RDF/XML, поскольку преобразование из формата MODS в RDF/XML можно осуществить стандартными средствами XML с использованием XSLT-шаблона. XSLT-шаблон приведен в Приложении Д.

3.2 Алгоритм связывания данных

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 библиотек России, недостаточно. Необходимо организовать связывание с уже имеющимися данными в LOD.

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

Поскольку в разработанной онтологии 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 широкоприменяемых 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 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 [90].

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, отсортированному 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 10.

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

Таблица 10: Синтаксисы запросов

Поскольку большая часть данных из Библиотеки конгресса США и Британской национальной библиотеки представлена на английском языке, а в России 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 [104].

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 отсеечения заведомо неподходящих кандидатов использовался метод биграмм, который применялся нами для выявления дублетных библиографических записей.

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. 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.

Очевидно, что количество уникальных значений «Автор» и «Название» меньше, чем количество записей. 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 базу 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.

Принципиальная схема работы системы поиска подобных записей и установления связей с ними приведена на рисунке 4. Пример RDF/XML файла со связями приведен в Приложении Д.

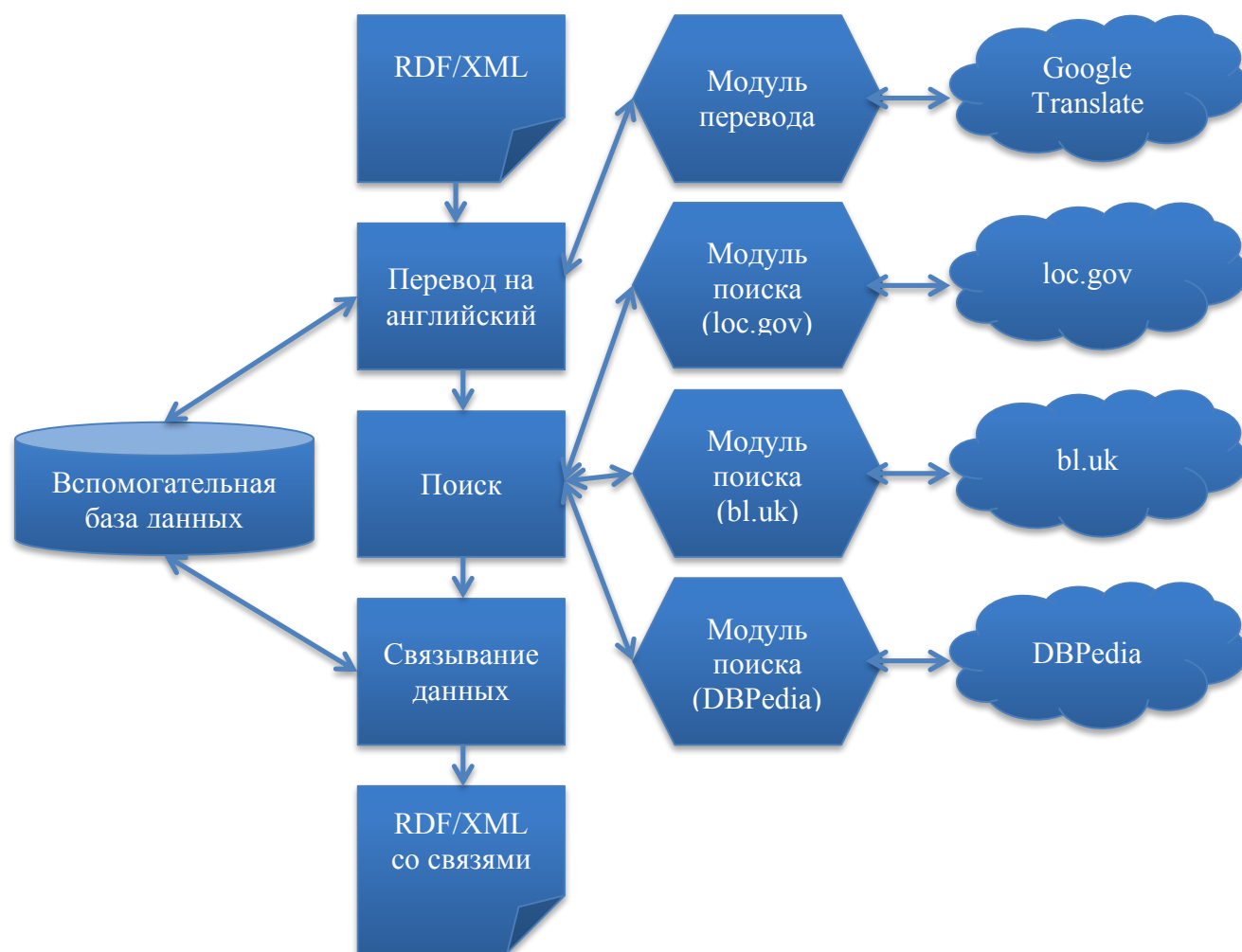


Рисунок 4: Архитектура системы семантического связывания данных

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.

3.3 Архитектура системы хранения данных и организации доступа

Научившись конвертировать библиографические записи в формат, пригодный для использования в семантической паутине, а также связывать их с другими данными в LOD, встает вопрос выбора способа хранения полученных RDF-триплетов, поскольку существуют различные механизмы хранения и предоставления доступа к данным, которые имеют как преимущества, так и недостатки. Сформулированные Тимом Бернерс-Ли основные принципы связанных данных [67] формально позволяют создать простейшее хранилище, состоящее из набора файлов и web-сервера. Данный web-сервер может предоставлять необходимую информацию в формате RDF при попытке доступа по URI. Очевидно, что такой подход не предоставляет возможность использовать SPARQL точку доступа, наличие которой стало стандартом де-факто.

Однако, вышеупомянутый подход плох не только отсутствием SPARQL точки доступа: при обращении к данным по URI должна происходить автоматическая конвертация файлов данных в формат RDF для каждого запроса. Подобная архитектура не является масштабируемой, поскольку при увеличении объема хранящейся информации и числа запросов производительность системы сильно деградирует.

В некоторых проектах хранение библиографических записей осуществляется в реляционной базе данных, а конвертация их в формат RDF [15] «на лету», что существенно экономит дисковое пространство за счет отсутствия дублирования, но при этом резко снижается производительность, и это снижение напрямую зависит как и от количества хранимых записей, так и от количества поступающих запросов.

Для решения вышеперечисленных проблем используют специализированные хранилища RDF-триплетов. Очевидно, что необходимость

хранения данных в сконвертированном виде приводит к дополнительному расходу пространства на сервере хранения. Однако, выгода от использования специализированных хранилищ гораздо больше: применение хранилищ RDF-триплетов позволяет достигнуть необходимого уровня масштабируемости, надежности, безопасности и быстродействия. Наиболее распространенными хранилищами являются 4store [146], Sesame [147] и TDB, входящим в состав интегрированной среды Jena [148].

Опыт использования хранилища 4store в проекте немецких научных библиотек Linking Open Bibliographic Data (LOBID) показал [78], что данное программное решение имеет ограниченную производительность: если запустить поиск на 700 миллионах RDF-триплетах, полученных из 16 миллионов библиографических записей, он будет выполняться недопустимо долго. Поскольку система семантической интеграции библиографических записей оперирует сопоставимыми объемами информации, использование RDF-хранилища 4store является неприемлемым.

По сравнению с Sesame TDB, входящая в состав Jena, обладает большей масштабируемостью и более быстрым откликом на запросы. В рейтинге баз данных она также занимает более высокое место [149]. Jena является быстроразвивающимся проектом, активно поддерживаемым огромным сообществом программистов. В состав этой среды входят компоненты, которые позволяют:

1. Хранить данные как в SQL базе данных, так и непосредственно в специализированном хранилище триплетов.
2. Использовать API для получения RDF-триплетов, информации об используемой онтологии, выполнения SPARQL-запросов.
3. С помощью конвертеров сериализовать полученные RDF-триплеты любым из способов: RDF/XML, Turtle, N-Triples, RDFa.
4. Предоставить RDF-данные и ответы на SPARQL-запросы с использованием протокола HTTP.

В состав Jena входит компонент, реализующий SPARQL точку доступа – Fuseki [150]. Эта система имеет тесную интеграцию с хранилищем триплетов TDB и представляет собой модуль, который способен обрабатывать запросы от внешних систем с использованием протокола SPARQL 1.1 для поиска и модификации связанных данных [151].

Для загрузки RDF-триплетов в хранилище TDB на языке Java был написан специальный модуль. Данный модуль использует Application Programming Interface (API) для языка Java, имеющийся в Jena. Эксперименты с передачей RDF-триплетов из базы данных в TDB показали, что обработка большого количества RDF/XML-файлов занимает значительно меньше времени, чем загрузка этих файлов по одному. На графике 1 приведена взаимосвязь скорости загрузки RDF-триплетов от количества триплетов, входящих в один пакет загрузки. Скорость измеряется количеством загруженных триплетов в секунду. Как видно из графика, оптимальным вариантом является загрузка RDF-триплетов пакетами по 400 тысяч. Дальнейшее увеличение количества RDF-триплетов в пакете не приводит к росту скорости загрузки, но оказывает негативное влияние на объем свободной оперативной памяти.

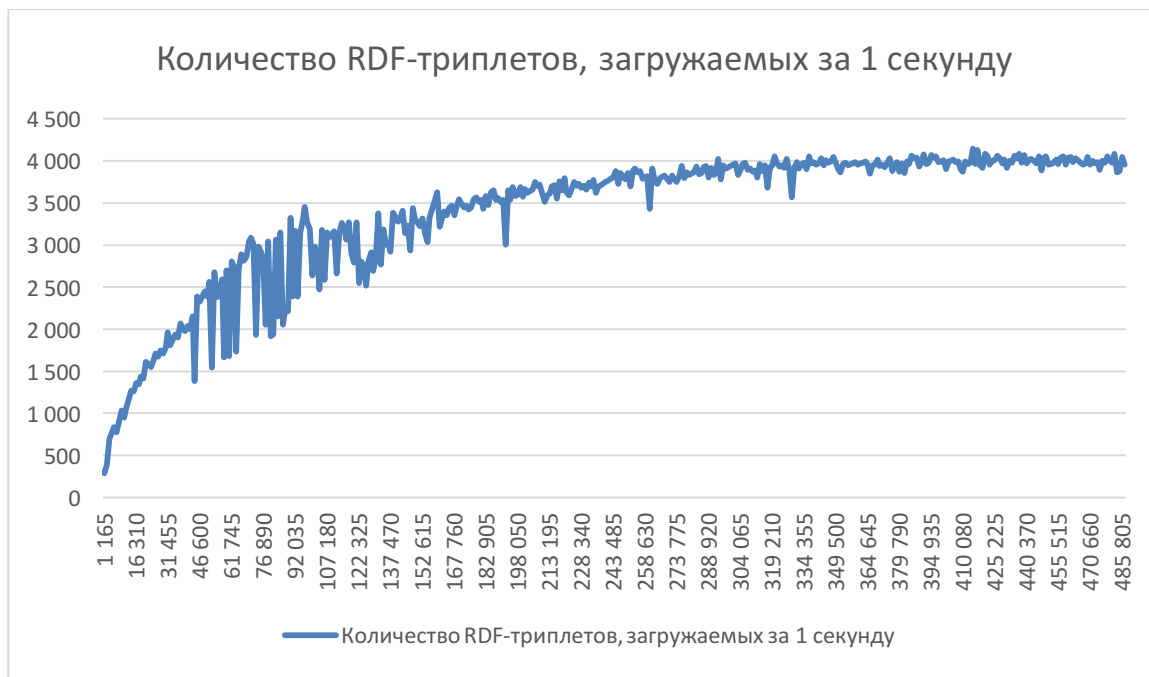


График 1: Скорость загрузки RDF-триплетов в хранилище TDB

Доступ к данным может осуществляться как по протоколу SPARQL, так и по протоколу HTTP. Для поддержки протокола HTTP был развернут и настроен веб-сервер с использованием Play Framework [152]. Данная среда позволяет создать высокопроизводительный веб-сервер с минимальным количеством настроек. Именно такой веб-сервер и требуется в нашей системе, поскольку его основной задачей является преобразование HTTP запросов в SPARQL с последующим обратным преобразованием ответов. При этом веб-сервер должен быть доступен максимальному количеству пользователей, что подразумевает высокую производительность и минимизацию механизмов безопасности.

Для интеграции с Fuseki используется механизм SOH – SPARQL Over HTTP, который представляет собой набор специализированных утилит – скриптов, входящих в состав Jena. Скрипты SOH написаны на языке Ruby, представляющим собой динамический, интерпретируемый, высокоуровневый язык программирования [87].

Реализованный с использованием Play Framework веб-сервер поддерживает два типа запросов: поиск и обновление данных. Для осуществления поиска используется скрипт s-query, который имеет синтаксис:

```
s-query --service=<URL SPARQL-сервера> 'query <запрос>'
```

А для обновления данных используется скрипт s-update, имеющий синтаксис:

```
s-update --service=<URL SPARQL-сервера> 'update <запрос>'
```

Результат работы скриптов проходит обратное преобразование и возвращается в качестве ответа на поступивший веб-серверу HTTP-запрос. На рисунке 5 показана архитектура системы интеграции данных.

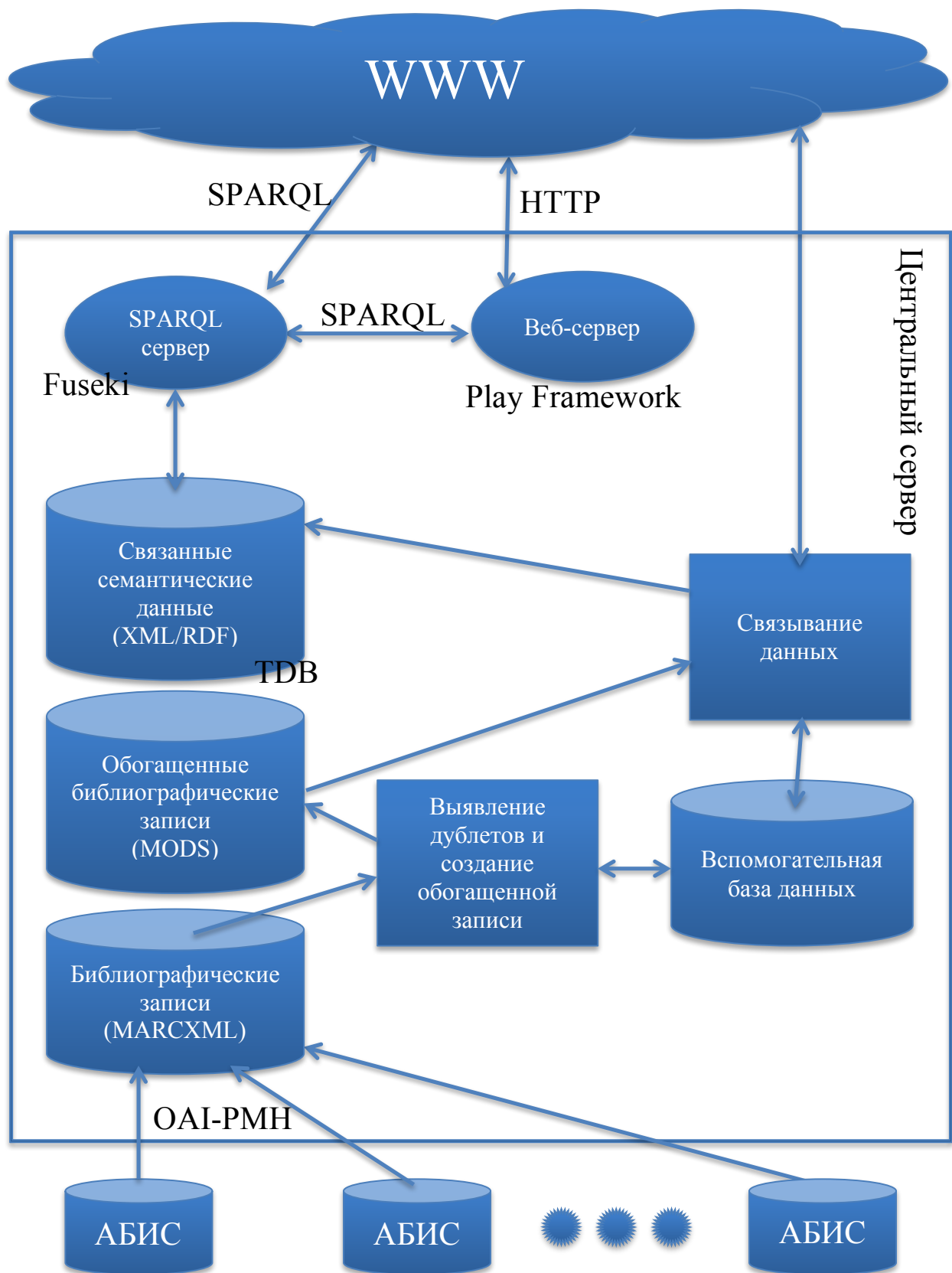


Рисунок 5: Архитектура системы интеграции данных

Наличие большого количества компонентов, поддерживаемых и развиваемых независимыми командами программистов, определило выбор в

пользу интегрированной среды Jena для хранения RDF-триплетов в системе семантической интеграции библиографических записей. Очевидно, что при росте количества хранимых триплетов производительность любого специализированного хранилища начнет деградировать. Возможным решением в данной ситуации является использование распределенных файловых систем, позволяющих распределить нагрузку между различными серверами [106].

Выводы главы 3

В данной главе приводится описание онтологии предметной области, совместимой с уже широкоиспользуемыми мировой общественностью словарями. Также приводится описание алгоритма поиска данных для связывания. Среди достоинств приведенного алгоритма можно выделить его универсальность, масштабируемость, возможность распараллеливания, использование механизмов кэширования. В заключительной части главы приводится подробное описание архитектуры системы интеграции библиографических данных, которая позволяет в автоматическом режиме осуществлять сбор записей, выявлять среди них дублетные, создавать на основе выявленных дублетов обогащенные записи, конвертировать созданные записи в формат, пригодный для публикации в LOD, находить подходящие данные для связывания и публиковать данные.

4 Апробация системы на электронном каталоге, консолидирующем библиографические записи из различных библиотек России

В данной главе приводится описание крупнейшего проекта по консолидации электронных документов в Российской Федерации – Национальной электронной библиотеки. Рассматривается структура и состав подсистем НЭБ. Приводятся результаты функционирования созданной системы интеграции библиографических данных на массиве записей электронного каталога Национальной электронной библиотеки.

4.1 Национальная электронная библиотека

В Министерстве культуры Российской Федерации предпринимаются попытки, направленные на реализацию нового этапа развития Национальной электронной библиотеки. 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 [141].

Достижение поставленной цели осуществляется путем решения ряда задач:

1. Формирование распределенного фонда, в состав которого входят 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.

Основной инфраструктурной особенностью функционирования НЭБ является одновременное наличие центрального портала и распределенного характера процессов создания, каталогизации, подключения, использования и хранения ресурсов среди участников НЭБ. Центральный портал обеспечивает навигацию и поиск по полным текстам и библиографическим записям распределенного фонда, доступ к ресурсам с учетом ограничений, накладываемых авторскими правами, функционирование единого электронного читательского билета, сбор статистики доступа к документам и составления отчетности, координацию процессов оцифровки различными библиотеками.

Основным отличием НЭБ от проектов подобного рода, например, сводного каталога электронных ресурсов (СКЭР) [142], является наличие единого интерфейса для работы с документами из различных библиотек. Предыдущие проекты предоставляли только ссылки на электронные ресурсы, для получения доступа к которым необходимо было регистрироваться в каждой конкретной библиотеке, устанавливать и настраивать специализированное программное обеспечение, применяемое в данной библиотеке.

Центральный портал НЭБ состоит из следующих подсистем:

- Подсистема управления распределенным фондом;
- Единый сводный каталог фонда НЭБ;
- Подсистема единого электронного читательского билета;
- Подсистема поиска;
- Подсистема просмотра полных текстов;
- Подсистема администрирования;
- Подсистема статистики и биллинга.

Подсистема управления распределенным фондом обладает следующим функционалом:

- Осуществляет сбор библиографических записей, содержащих данные об экземпляре издания, его формате и местоположении, для формирования единого сводного каталога НЭБ.
- Осуществляет индексацию распределенного фонда с последующей интеграцией индекса в подсистему поиска.

Единый сводный каталог фонда НЭБ обеспечивает:

- Навигацию по распределенному фонду НЭБ.
- Структурирование и систематизацию библиографической, справочной и адресной информации.
- Отображение информации об издании, в том числе и сведения об ограничениях использования, накладываемых правообладателями.
- Координацию усилий библиотек по оптимизации и планированию процессов оцифровки с целью исключения многократного сканирования одних и тех же ресурсов в разных библиотеках.

Подсистема единого электронного читательского билета обладает следующим функционалом:

- Создание и активация единого электронного читательского билета через центральный портал НЭБ.
- Аутентификация и авторизация пользователей с использованием единого электронного читательского билета.

Подсистема поиска реализует возможность:

- Полнотекстового поиска по всему массиву текстов электронных изданий.
- Библиографического поиска по единому сводному каталогу фонда НЭБ.
- Использования системы фильтров и запросов, позволяющих сформулировать набор критериев для узконаправленного, специализированного, уточняющего поиска.

Подсистема просмотра полных текстов позволяет:

- Просмотреть содержимое любых документов, находящихся в фонде НЭБ, как с использованием веб-приложения, так и с помощью мобильного приложения.
- Осуществить поиск внутри просматриваемых документов.
- Пометить весь документ произвольным тэгом.
- Выделить часть документа для последующего цитирования.

Подсистема администрирования обеспечивает:

- Единую точку доступа через центральный портал к фонду НЭБ.
- Разграничение прав доступа между различными категориями пользователей.
- Защиту информационной системы от несанкционированного доступа.
- Регистрацию, аутентификацию и учет пользователей.
- Мониторинг функциональности системы для предотвращения аварийных ситуаций.

Подсистема статистики и биллинга отвечает за:

- Сбор, хранение и обработку данных об использовании фондов НЭБ, а также других действий пользователей.
- Контроль использования изданий с учетом ограничений, накладываемых правообладателем.
- Функционирование возмездных сервисов.
- Формирование и использование реестра правообладателей, содержащего полную информацию о действующих правах на каждое издание в фонде НЭБ.

На сентябрь 2015 года в НЭБ представлены ресурсы 6 федеральных, 29 региональных библиотек и 1 библиотеки ВУЗа, с помощью единого каталога пользователь имеет возможность получить информацию о местонахождении более 33 миллионов книг в этих библиотеках. Электронный каталог НЭБ представляет собой уникальный по своей полноте массив библиографических записей из библиотек различной ведомственной подчиненности, использующих

разные технологические решения для создания, хранения и обеспечения доступа к метаданным. Для интеграции библиографических записей в единый каталог на центральном портале необходимо выявить дублетные записи и осуществить их слияние.

4.2 Результаты работы алгоритма выявления дублетных библиографических записей

Алгоритм выявления дублетных библиографических записей выполняется на центральном сервере НЭБ, который по данным программы синтетических тестов AIDA64 [143] обладает следующими характеристиками:

1. Процессор - двухпроцессорный Intel(R) Xeon(R) CPU E5645 @ 2.40 GHz, обладающий 12 реальными и 24 виртуальными ядрами.
2. Жёсткий диск – RAID0 Intel Multi-Flex SCSI Disk Device. Средняя скорость чтения данных – 150 MB/s.
3. Оперативная память – 98 238 MB DDR3 1066 MHz.

На сервере установлена операционная система Microsoft Windows Server 2012 Standard, а в качестве СУБД используется Microsoft SQL Server 2012 – 11.0.2100.60 (X64).

Алгоритм выявления дублетных библиографических записей обработал 21 313 009 записей. Для обработки такого объема данных на указанном сервере потребовалось 90 часов. Столь небольшого времени работы алгоритма удалось добиться за счет того, что вся хранимая в базе данных информация, занимающая всего 60 гигабайт, была полностью подгружена в оперативную память. Это позволило избавиться от относительно медленных операций чтения/записи из/в СУБД.

В результате работы алгоритма было выявлено, что 1 239 293 библиографические записи являются дублетами какой-нибудь другой записи. Уникальных записей, т.е. записей, для которых не существует ни одного дублета,

оказалось 19 066 057. Статистика по количеству дублетных записей записей приведена в таблице 11.

Количество дублетных записей	Количество записей
0	19 066 057
1	699 687
2	101 490
3	29 247
4	11 655
5	6 777
...	
141	1
151	1
153	1

Таблица 11: Количество записей с определенным количеством дублетов.

Как видно из вышеприведенной таблицы, наибольшее количество записей не имеют дублетов, однако, имеется запись, для которой было найдено 153 дублетных записи.

Как оказалось, заполненное поле ISBN имеют далеко не все книги, поэтому на этапе детерминированного поиска было выявлено только 0,6% дублетных записей. Алгоритм разбиения всех хешей на кластеры отработал таким образом, что в среднем в одном кластере получилось по 941 записи. Оптимизация по полю «Год издания» позволила не считать меры Жаккара в среднем для 71,4% пар записей из кластера. В среднем одна библиографическая запись участвовала в качестве аргумента для подсчета меры Жаккара 264 раза. С помощью использования Теоремы №1 удалось не считать меру Жаккара для 0,103% пар записей, использование Теоремы №2 позволило не досчитывать до конца меру Жаккара для 0,013% пар записей. Столь низкий процент срабатывания теорем, скорее всего, объясняется достаточно эффективной работой алгоритма по разбиению на кластеры.

4.3 Публикация библиографических записей НЭБ в Linked Open Data

После создания обогащенных библиографических записей к полученному массиву был применен алгоритм поиска и связывания с уже опубликованными данными. Оказалось, что среди обрабатываемого массива библиографических записей количество уникальных значений поля «Автор» составляет 19,6% от общего числа записей. Для поля «Название» данное значение составляет 68,6%. Таким образом, кэширование результатов запросов к онлайн-сервисам позволило уменьшить количество запросов 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 [145] 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. Информация о скорости соединения с интернетом приведена на рисунке 6. 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 миллионов, 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 которых обрабатывает независимое множество RDF/XML-файлов. Подобная оптимизация позволила сократить время работы алгоритма до 30 часов.

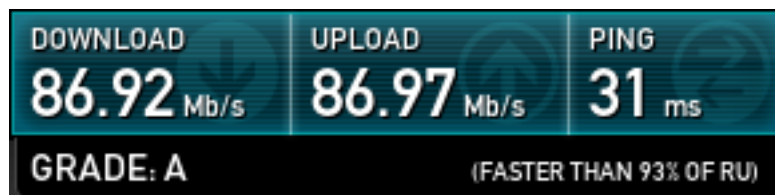


Рисунок 6: Скорость соединения центрального сервера НЭБ с интернетом

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 приведены в таблице 12.

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

Таблица 12: Среднее количество RDF-триплетов связи в одном RDF/XML-файле

На рисунке 7 показан результат обращения к веб-серверу по URI – RDF-файл, сериализованный с использованием Turtle.

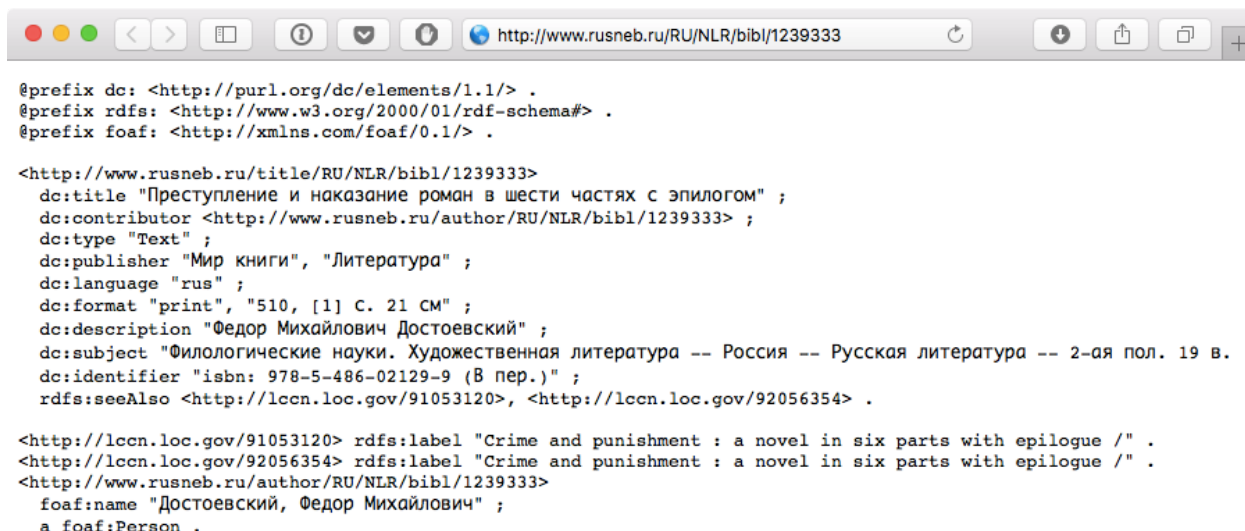


Рисунок 7: Результат обращения к веб-серверу

Выводы главы 4

Национальная электронная библиотека – проект, призванный стать единым общим национальным электронным пространством знаний. Данный проект является крупнейшим хранилищем библиографических записей, когда-либо созданных на территории России. Методы и алгоритмы интеграции, представленные в диссертационном исследовании, были успешно применены на электронном каталоге НЭБ, доказав на практическом опыте свою универсальность и масштабируемость. Уникальность разработанных подходов была доказана тем, что созданная система интеграции смогла обработать крупнейший массив библиографических данных за рекордно короткие сроки.

Заключение

В диссертационной работе выполнен анализ существующих и перспективных подходов, технологий, онтологий, алгоритмов и систем, применяемых для консолидации огромного количества библиографических записей из разнородных источников с последующей публикацией данных с использованием технологий семантической паутины. Основной целью работы является расширение возможностей интеграции библиографических записей в открытое семантическое пространство. Для достижения этой цели была поставлена и успешно решена научная задача по разработке методов и алгоритмов интеграции библиографических записей в открытое семантическое пространство. Создана модульная система, осуществляющая сбор библиографических записей, конвертацию собранных данных и публикацию их в LOD. Реализованы модули сбора библиографических записей из различных автоматизированных библиотечных интегрированных систем. Разработан и реализован алгоритм выявления дублетных библиографических записей. Реализован механизм создания обогащенной библиографической записи на основе выявленных дублетов. Созданы XSLT-шаблоны для конвертации полученных данных в формат, согласующийся с широко используемыми в библиотечном мире онтологиями представления данных. Разработан и реализован алгоритм выявления данных в LOD, подходящих для связывания. Создан программный комплекс, позволяющий хранить, обрабатывать и предоставлять доступ к данным с использованием протоколов HTTP и SPARQL. Разработанная система интеграции успешно прошла апробацию на электронном каталоге крупнейшего проекта по интеграции библиографических записей в России – Национальной электронной библиотеке.

Основными научными и практическими результатами проделанной диссертационной работы являются:

1. Структура и формат онтологии для публикации данных, полученных из библиографических записей, в открытом семантическом пространстве.

2. Масштабируемый алгоритм установления близости текстовых полей в библиографических записях на основе разбиения текстовых значений на биграммы с последующим отсечением с использованием меры Жаккара.
3. Алгоритм обнаружения дублетных библиографических записей и аддитивного пополнения данных в LOD в автоматическом режиме с использованием разработанной онтологии предметной области.
4. Архитектура сбора, хранения и публикации библиографических данных, позволяющая в автоматическом режиме осуществлять сбор библиографических записей из различных библиотек России, конвертировать их в формат, пригодный для публикации в LOD, проводить аддитивное пополнение данных и устанавливать связи с уже опубликованными в LOD данными.

Список сокращений и условных обозначений

API – Application Programming Interface

DCMI – Dublin Core Metadata Initiative

DOM – Document Object Model

EDM – Europeana Data Model

ESE – Europeana's Semantic Elements

FOAF – Friend of a Friend

HTTP – Hypertext Transfer Protocol

ISBN – International Standard Book Number

JSON – JavaScript Object Notation

LOBID – Linking Open Bibliographic Data

LOD – Linked Open Data

MADS – Metadata Authority Description Schema

MARC – Machine-Readable Cataloging

MODS – Metadata Object Description Schema

MVC – Model-View-Controller

OAI-PMH – Open Archives Initiative Protocol for Metadata Harvesting

OWL – Web Ontology Language

RDF – Resource Description Framework

RDFS – Resource Description Framework Schema

SKOS – Simple Knowledge Organization System

SOH – SPARQL Over HTTP

SPARQL – SPARQL Protocol And RDF Query Language

URI – Uniform Resource Identifier

URL – Uniform Resource Locator

XML – Extensible Markup Language

XSLT – Extensible Stylesheet Language Transformation

АБИС – автоматизированная библиотечная интегрированная система

ЕНИП РАН – Единое научное информационное пространство Российской академии наук

НЭБ – Национальная электронная библиотека

РГБ – Российская государственная библиотека

РНБ – Российская национальная библиотека

СКБР – Сводный каталог библиотек России

СКЭР – Сводный каталог электронных ресурсов

СУБД – Система управления базами данных

Список литературы

1. Принципы разработки интегрированной системы для научных библиотек, архивов и музеев / А. Б. Антопольский, А. А. Каленкова, Н. Е. Каленов [и др.] // Информ. ресурсы России. — 2012. — № 1. — С. 2—6.
2. Атаева, О. М. Методы очистки интегрируемых данных / О. М. Атаева, Л. Н. Шиолашвили // Труды 49-й научной конференции МФТИ «Современные проблемы фундаментальных и прикладных наук» (Москва – Долгопрудный, 24 – 25 нояб. 2006 г.). — Москва ; Долгопрудный, 2006. — Ч. VII : Управление и прикладная математика. — С. 243—244.
3. Бахтурина, Т. А. Структура и объект библиографической записи. Пунктуация / Т. А. Бахтурина // Библиотека. — 2004. — № 3. — С. 48–50.
4. Бахтурина, Т. А. Будущее каталогизации в России и в мире / Т. А. Бахтурина // Науч. и техн. б-ки. — 2010. — № 9. — С. 34—44.
5. Библиографическое описание электронных ресурсов : метод. рекомендации / [сост.: Т. А. Бахтурина и др.] ; Рос. гос. б-ка, Межрегион. ком. по каталогизации. — Москва : Пашков дом, 2002. — 55 с.
6. Интеграция метаданных единого научного информационного пространства РАН / А. А. Бездушный, А. Н. Бездушный, В. А. Серебряков, В. И. Филиппов ; Рос. акад. наук, Межведомств. суперкомпьютер. центр, Вычисл. центр им. А. А. Дородницына. — Москва : Вычисл. Центр РАН (ВЦ РАН), 2006. — 238 с. : ил.
7. Диссертационные материалы в онлайн-библиотеке / В. М. Богданова, М. М. Горбунов-Посадов, Т. А. Полилова, М. И. Слепенков // Препринты ИПМ им. М. В. Келдыша РАН. — 2015. — № 62. — 12 с.
8. Борисенко, Т. В. Правила библиографического описания произведений печати для массовых библиотек / [сост.: Т. В. Борисенко, И. С. Дудник] ; Гос. б-ка СССР им. В. И. Ленина. — Москва : Б.и., 1978. — 50 с.
9. Воройский, Ф. С. Основы проектирования автоматизированных библиотечно-информационных систем / Ф. С. Воройский. — Москва : Физматлит, 2002. — 383 с. : ил.

10. Воройский, Ф. С. Корпоративные автоматизированные библиотечно-информационные системы – классификация и принципы построения / Ф. С. Воройский, Я. Л. Шрайберг [Электронный ресурс] // Электрон. б-ки : электрон. науч. журн. — Электрон. дан. — 2002. — Т. 5, вып. 5. — Режим доступа: <http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2002/part5/SV>, свободный. — Загл. с экрана.
11. Воройский, Ф. С. Региональные корпоративные библиотечные системы России: опыт последних трех лет глазами участников движения / Ф. С. Воройский, Я. Л. Шрайберг // Науч. и техн. б-ки. — 2003. — № 3. — С. 13—33.
12. Гиляревский, Р. С. Развитие принципов книгоописания : крат. очерк / Р. С. Гиляревский. — Изд. 2-е, доп. — Санкт-Петербург : Профессия, 2008. — 236, [2] с. — (Библиотека).
13. Жабко, Е. Д. Справочно-библиографическое обслуживание в электронной среде: теория и практика : монография / Е. Д. Жабко ; Рос. нац. б-ка. — Санкт-Петербург : Рос. нац. б-ка, 2006. — 387 с.
14. Жлобинская, О. Н. Semantic Web, связанные данные и библиотеки : докл. на конф. «ЛИБНЕТ-2012», г. Звенигород, 12 – 23 нояб. 2012 г. [Электронный ресурс] / О. Н. Жлобинская // Российский коммуникативный формат : офиц. сайт. — Электрон. дан. — [Москва, 2015]. — 31 с. — Режим доступа: <http://www.rusmarc.ru/publish/SemWeb.pdf>, свободный. — Загл. с экрана.
15. Жлобинская, О. Н. Как представлять библиографические данные в RDF: докл. на конф. «ЛИБНЕТ-2013», г. Звенигород, 18 – 22 нояб. 2013 г. [Электронный ресурс] / О. Н. Жлобинская // Российский коммуникативный формат : офиц. сайт. — Электрон. дан. — [Москва, 2015]. — 11 с. — Режим доступа: <http://www.rusmarc.ru/publish/bibdateRDF.pdf>, свободный. — Загл. с экрана.
16. Методические рекомендации по каталогизации и индексированию электронных копий документов / [Л. В. Завьялова и др.; под общ. ред. Е. Д. Жабко] ; Президент. б-ка. — Санкт-Петербург : Президент. б-ка, 2014. — 287 с.

17. Предметный поиск и машиночитаемая каталогизация : [сб. ст.] / сост. С. Б. Голубцов ; ред.: Е. И. Загорская, Н. П. Никольцева ; Рос. нац. б-ка. — Санкт-Петербург : Рос. нац. б-ка, 2014. — Вып. 1. — 183, [1] с.
18. Земсков, А. И. Электронные библиотеки : учеб. для вузов / А. И. Земсков, Я. Л. Шрайберг. — Москва : Либерия, 2003. — 351 с. : ил. — (Альманах «Приложение к журналу «Библиотека» ; 2-е полугодие 2003 г.).
19. Каленов, Н. Е. Электронная библиотека «Научное наследие России» / Н. Е. Каленов, Г. И. Савин, А. Н. Сотников // Информ. ресурсы России. — 2009. — № 2. — С. 19—20.
20. Каленов, Н. Е. Технология создания электронной библиотеки «Научное наследие России» / Н. Е. Каленов, Г. И. Савин, А. Н. Сотников // Научная книга на постсоветском пространстве : материалы II Междунар. конф. (г. Москва, 19 – 21 сент. 2007 г.) / Междунар. ассоц. акад. наук. — Москва : Наука, 2007. — С. 11—16.
21. Калинина, Г. П. Формирование заголовка библиографической записи. Имя лица в заголовке / Г. П. Калинина // Библиотека. — 2004. — № 8. — С. 37—42.
22. Калинина, Г. П. Составление стандартной библиографической записи на книги : практ. рекомендации / Г. П. Калинина ; Рос. кн. палата. — Москва : РКП, 2006. — 167 с. — (Библиотека Российской книжной палаты : методические материалы и рекомендации ; вып. 2).
23. Каспарова, Н. Н. Особенности состава данных и структуры библиографической записи в электронной среде / Н. Н. Каспарова // Библиотековедение. — 2008. — № 2. — С. 51—56.
24. Российские правила каталогизации : [в 2 ч. / авт. коллектив: Н. Н. Каспарова (рук.) и др. ; редкол.: Н. Н. Каспарова (гл. ред.) и др.] ; Рос. библ. ассоц., Межрегион. ком. по каталогизации]. — Москва : Пашков дом, 2008. — Ч. 1 : Основные положения и правила. — 242, [1] с.
25. Каспарова, Н. Н. Российские правила каталогизации : [в 2 ч. / авт. коллектив: Н. Н. Каспарова (рук.) и др. ; редкол.: Н. Н. Каспарова (гл. ред.) и др.] ; Рос. библ.

- ассоц., Межрегион. ком. по каталогизации]. — Москва : Пашков дом, 2008. — Ч. 2 : Специальные правила каталогизации отдельных видов документов. — 419 с.
26. Князева, А. А. Автоматическое связывание записей библиографических баз данных на основе унифицированных поисковых признаков : дис. ... канд. техн. наук : 05.25.05 / Князева Анна Анатольевна ; Ин-т вычисл. технологий СО РАН. — Новосибирск, 2014. — 147 с. : ил.
27. Князева, А. А. Выявление дубликатов в библиографических базах данных [Электронный ресурс] / А. А. Князева, И. Ю. Турчановский, О. С. Колобов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции : [сб. тр. XV Всерос. науч. конф. RCDL 2013, г. Ярославль, 14 – 17 окт. 2013 г.]. — Электрон. дан. — Ярославль: ЯрГУ им. П. Г. Демидова, 2013. — С. 276—282. — Режим доступа: http://rcdl.ru/doc/2013/paper/s10_4.pdf, свободный. — Загл. с экрана.
28. Колосов, К. А. Создание и применение в библиотечной практике корпоративной технологии на базе протокола Z39.50 : автореф. дис. ... канд. техн. наук : 05.25.05 / Колосов Кирилл Анатольевич ; [Моск. гос. ун-т культуры и искусств]. — Москва, 2008. — 18 с.
29. Колосов, К. А. Решения Системы автоматизации библиотек ИРБИС для региональных порталов государственных и муниципальных услуг / К. А. Колосов // Науч. и техн. б-ки. — 2012. — № 11. — С. 61—66.
30. Ленской, Б. В. Книжный рынок России. Состояние, тенденции и перспективы развития : отраслевой докл. / Федер. агентство по печати и массовым коммуникациям, Упр. период. печати, книгоиздания и полиграфии; [подгот. авт. коллективом в составе Б. В. Ленского и др. ; под общ. ред. В. В. Григорьева]. — Москва : Федер. агентство по печати и массовым коммуникациям, 2015. — 112 с.
31. Логинов, Б. Р. СКБР – общероссийский навигатор библиотечных ресурсов и услуг : [презентация выступления, представл. на пленар. заседании Ежегод. совещ. рук. федер. и центр. б-к субъектов Рос. Федерации на тему: «Формирование и использование электронных ресурсов библиотек России», г. Москва, 23 окт. 2012 г.] / Б. Р. Логинов // ГосБук: эксперт. сеть по вопр. гос. упр.

— Электрон. дан. — [20 нояб. 2012 г.]. — 18 с. — Режим доступа: <http://www.gosbook.ru/document/64912/64947/preview>, свободный. — Загл. с экрана.

32. Логинов, Б. Р. Чем сложнее задача, тем интереснее решение / Б. Р. Логинов // Современ. б-ка. — 2009. — № 2. — С. 11—17.
33. Маршак, Б. И. Направления развития программного обеспечения и технического комплекса при автоматизации библиотечно-информационных процессов ГПНТБ России / Б. И. Маршак // Науч. и техн. б-ки. — 1997. — № 1. — С. 69—72.
34. Маршак, Б. И. Проблемы и решения в области автоматизации в корпоративной телекоммуникационной библиотечной среде / Б. И. Маршак // Науч. и техн. б-ки. — 2001. — № 3. — С. 42—44.
35. Масхулия, Т. Л. Заголовок, содержащий наименование организации / Т. Л. Масхулия // Библиотека. — 2004. — № 9. — С. 31—32.
36. Немировский, Е. Л. Изобретение Иоганна Гуттенберга = Johann Gutenberg's invention: из истории книгопечатания. Технические аспекты / Е. Л. Немировский ; под ред. В. И. Васильева. — Москва : Наука, 2000. — 658, [1] с. : ил., портр.
37. [Оленин, А. Н.] Опыт нового библиографического порядка для Санкт-Петербургской имп. библиотеки : представлен от д. с. с. Оленина и утв. гл. начальством в 1808 г / [А. Н. Оленин]. — Санкт-Петербург : тип. Губ. правл., 1809. — [8], 112 с.
38. Племнек, А. И. Концепция программы «Автоматизация библиотек» / А. И. Племнек // Библиотека. — 1999. — № 10. — С. 32—35.
39. Племнек, А. И. Центр «Открытые Библиотечные Системы»: опыт разработки корпоративной библиотечной системы регионального уровня / А. И. Племнек, Н. В. Соколова // Информ. бюл. Рос. библ. ассоц. — 1999. — № 15. — С. 246—255.
40. Рубцов, Д. Н. Выявление дубликатов в разнородных библиографических источниках / Д. Н. Рубцов, В. Б. Барахнин // Вестн. Новосиб. гос. ун-та. Серия : Информ. технологии. — 2009. — Т. 7, № 3. — С. 86—93.

41. Селиванова, Ю. Г. Каталогизация : современные технологии. Тенденции и перспективы развития / Ю. Г. Селиванова, Т. Л. Масхулия, О. Н. Жлобинская [и др.] ; Нац. информ.-библ. центр ЛИБНЕТ, Рос. нац. б-ка, Рос. библ. ассоц. — Москва : Фаир [и др.], 2007. — 215 с. : ил.
42. Скворцов, В. В. Форматы и стандарты библиографических данных в информатизации библиотечного дела / В. В. Скворцов // Библ. компьютер. сети. — 2007. — С. 57—65.
43. Скворцов, В. В. Программа представления библиографических данных в Семантическом Вебе : докл. на конф. «ЛИБНЕТ-2013», г. Звенигород, 18 – 22 нояб. 2013 г. [Электронный ресурс] / В. В. Скворцов // Российский коммуникативный формат : офиц. сайт. — Электрон. дан. — [Москва, 2015]. — 9 с. — Режим доступа: http://www.rusmarc.ru/publish/rdf_doklad.pdf, свободный. — Загл. с экрана.
44. Скрипкина, Т. И. Библиотека Оксфордского университета / Т. И. Скрипкина // Тр. Б-ки Акад. наук СССР и Фундам. б-ки обществ. наук АН СССР. — 1959. — Т. 4. — С. 288—295.
45. Соболящikov, В. И. Об устройстве общественных библиотек и составление их каталогов / [соч.] Василия Соболящикова, ст. библиотекаря Имп. Публ. б-ки. — Санкт-Петербург: тип. Имп. Акад. наук, 1859. — [6], 56 с.
46. Сысойкина, М. А. Протокол Z39.50: обоснование необходимости использования [Электронный ресурс] / М. А. Сысойкина // Информационные технологии в книжном бизнесе. — Электрон. дан. — [] Режим доступа: <http://www.bookresearch.ru/Z3950usage.htm>, свободный. — Загл. с экрана.
47. Шорин, О. Н. Алгоритм слияния дублетных библиографических записей / К. А. Косолапов, В. А. Серебряков, К. Б. Теймуразов, О. Н. Шорин // Научный сервис в сети Интернет: труды XVII Всероссийской научной конференции (21-26 сентября 2015 г., г. Новороссийск). — М.: ИПМ им. М. В. Келдыша, 2015. — 336 с. — С. 173-181.
48. Шорин, О. Н. Интеграция библиографических записей / О. Н. Шорин // Информационные ресурсы России. — 2016. — №5. — С. 14-18.

49. Шорин, О. Н. Методы и средства интеграции и обогащения библиотечных данных. / К. А. Косолапов, В. А. Серебряков, К. Б. Теймуразов, О. Н. Шорин // Сборник избранных научных статей. Труды Пятого Всероссийского симпозиума «Инфраструктура научных информационных ресурсов и систем» (С.-Петербург, 6–8 октября 2015 г.).— Под ред. Е. В. Кудашева, В. А. Серебрякова.— В 2-х тт.— М.: ВЦ РАН, 2015.— Т.1.— С. 152-164.
50. Шорин, О. Н. Национальная электронная библиотека: технологии реализации / О. Н. Шорин // Национальная библиотека.— 2015.— №3(06).— С. 60-65.
51. Шорин, О. Н. Проблемы семантической интеграции библиотечных данных / В. А. Серебряков, О. Н. Шорин // Библиотековедение.— 2014.— №5.— С. 41-47.
52. Шорин, О.Н. Публикация библиографических данных в открытое семантическое пространство / О. Н. Шорин // Информационные ресурсы России.— 2016.— №4.— С. 19-23.
53. Шорин, О. Н. Сбор, обработка и хранение библиографических записей с использованием технологий семантической паутины / О. Н. Шорин // Библиотековедение.— 2015.— №2.— С. 37-42.
54. Шорин, О. Н. Семантическая интеграция библиотечных данных / В. А. Серебряков, К. Б. Теймуразов, О. Н. Шорин // Сборник избранных научных статей. Труды Четвертого Всероссийского симпозиума «Инфраструктура научных информационных ресурсов и систем» (С.-Петербург, 6–8 октября 2014 г.).— Под ред. Е. В. Кудашева, В. А. Серебрякова.— В 2-х тт.— М.: ВЦ РАН, 2014.— Т.1.— С. 83-96.
55. Шорин, О. Н. Современные тенденции поиска и анализа информации / О. Н. Шорин // Национальная библиотека.— 2015.— №1(04).— С. 34-39.
56. Шорин, О. Н. Стратегические IT инновации в Российской национальной библиотеке / А. В. Лихоманов, О. Н. Шорин // Медиатека и мир.— 2014.— №3.— С. 4-8.
57. Шрайберг, Я. Л. Основные положения и принципы разработки автоматизированных библиотечно-информационных систем и сетей: главные

тенденции окружения, основные положения и предпосылки, базовые принципы : учебно-практ. пособие / Я. Л. Шрайберг. — 2-е изд., испр. и доп. — Москва : Либерия, 2001. — 104 с. : ил. — (Альманах «Приложение к журналу «Библиотека» ; 2-е полугодие 2001 г.).

58. Шрайберг, Я. Л. Автоматизированные библиотечно-информационные системы России: состояние, выбор, внедрение, развитие = Automated library-information systems of Russia: current state, selection, implementation, development / Я. Л. Шрайберг, Ф. С. Воройский. — Москва : Либерия, 1996. — 270, [1] с. : ил.

59. Экстрем, М. В. Унифицированное заглавие / М. В. Экстрем, Т. А. Паршина, Г. Р. Ридер // Библ. дело. — 2003. — № 4. — С. 18—20.

60. Alemu, G. Linked Data for libraries: benefits of a conceptual shift from library-specific record structures to RDF-based data models / Getaneh Alemu, Brett Stevens, Penny Ross, Jane Chandler // New Libr. World. — 2012. — Т. 113, N 11-12. — P. 549—570.

61. 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.

62. Ausbrooks, R. Mathematical Markup Language (MathML): version 3.0 : W3C Recommendation 21 Oct. 2010 [Electronic resource] / Ron Ausbrooks, Stephen Buswell, David Carlisle [et al.] ; ed. by: David Carlisle, Patrick Ion, Robert Miner // W3C: World Wide Web Consortium. — Electronic data. — [Keio, Korea, Jap. ; Cambridge, MA, USA ; Biot, France] ; Beihang, China, 1998—2010. — 385 p. — Mode of access: <https://www.w3.org/TR/2010/REC-MathML3-20101021/mathml.pdf>, free. — Title of screen.

63. OAI for beginners – the Open Archives Forum online tutorial. Pt. 2: History and development of OAI-PMH [Electronic resource] // Open Archives Forum / Inform. Soc. Technologies. — Electronic data. — [Brussels, Belgium], [2003]. — P. 2. — Mode of access: <https://www.oaforum.org/tutorial/english/page2.htm>, free. — Title of screen.

64. Beckett, D. Turtle – terse RDF triple language : W3C Team Submission 28 Mar. 2011 [Электронный ресурс] / David Beckett, Tim Berners-Lee // W3C: World Wide

Web Consortium. — Electronic data. — Keio, Korea, Jap. ; Cambridge, MA, USA ; Biot, France, 2008. — Mode of access: <https://www.w3.org/TeamSubmission/turtle/>, free. — Title of screen.

65. Beckett, D. RDF/XML syntax specification: (rev.) : W3C Working Draft 25 Mar. 2002 [Electronic resource] / ed. by Dave Beckett // W3C: World Wide Web Consortium. — Electronic data. — Keio, Korea, Jap. ; Cambridge, MA, USA ; Le Chesnay, France, 2002. — Mode of access: <https://www.w3.org/TR/2002/WD-rdf-syntax-grammar-20020325/>, free. — Title of screen.

66. Bermes, E. Le web sémantique en bibliothèque / Emmanuelle Bermes ; avec la collab. d'Antoine Isaac et Gautier Poupeau. — Paris : Electre ; Cercle de la libr., 2013. — 171 p. : ill.

67. Berners-Lee, T. Linked Data. [Design issues] [Electronic resource] / 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.

68. Berners-Lee, T. Giant Global Graph [Electronic resource] / Tim Berners-Lee // Decentralized Information Group (DIG) Breadcrumbs. — Electronic data. — [Cambridge, MA, USA]. — 21/11/2007. — Mode of access: <http://dig.csail.mit.edu/breadcrumbs/node/215>, free. — Title of screen.

69. Berners-Lee, T. Uniform Resource Identifiers (URI): generic syntax [Electronic resource] / Tim Berners-Lee, Roy T. Fielding, Larry Masinter // IETF: Internet Engineering Task Force / Internet Soc. — Electronic data. — [Fremont, CA, USA], 1998. — RFC 2396. — 40 p. — Mode of access: <http://tools.ietf.org/pdf/rfc2396.pdf>, free. — Title of screen.

70. Berners-Lee, T. The Semantic Web / Tim Berners-Lee, James Hendler, Ora Lassila // Sci. Amer. — 2001. — N 5 (May). — P. 29—37.

71. Bizer, C. Linked Data – the story so far / Christian Bizer, Tom Heath, Tim Berners-Lee // IJSWIS: Int. j. on Semantic Web a. inform. systems. — 2009. — Vol. 5, N 3. — P. 1—22.

72. Böhme, C. Towards an infrastructure for the synchronisation of metadata in library [Electronic resource] / Christoph Böhme // SWIB'12: Semantic Web in libraries, Nov. 26 – 28, 2012, Cologne, Germany. — Electronic data. — 28 Nov. 2012. — Abstr., slides [14 p.], video. — Mode of access: <http://swib.org/swib12/programme.php>, free. — Title of screen.
73. Kernighan, B. W. The C programming language / Brian W. Kernighan, Dennis M. Ritchie. — 2nd ed. — Englewood Cliffs, N.J. : Prentice Hall, 1988. — XII, 272 p.
74. Brickley, D. RDF Schema 1.1 : W3C Recommendation 25 Febr. 2014 [Electronic resource] / Dan Brickley, Ramanathan V. Guha // 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/rdf-schema/>, free. — Title of screen.
75. Brickley, D. FOAF Vocabulary Specification 0.99 : Namespace doc. 14 Jan. 2014 – Paddington ed. [Electronic resource] / Dan Brickley, Libbi Miller ; Friend of a Friend (FOAF) : an experimental linked inform. system. — Electronic data. — 2000—2014. — Mode of access: <http://xmlns.com/foaf/spec/>, free. — Title of screen.
76. 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).
77. Charikar, M. S. Similarity estimation techniques from rounding algorithms / Moses S. Charikar // Proceedings of the forty-fourth annual ACM symposium on Theory of computing – STOC'02, Montreal, QC, Canada, May 19 – 21, 2002. — New York : ACM, 2002. — P. 380—388.
78. Christoph, P. Building a high performance environment for RDF publishing [Electronic resource] / Pascal Christoph // SWIB'12: Semantic Web in libraries, Nov. 26 – 28, 2012, Cologne, Germany. — Electronic data. — 27 Nov. 2012. — Abstr., slides [85 p.], video. — Mode of access: <http://swib.org/swib12/programme.php>, free. — Title of screen.

79. Davison, S. Enhancing an OAI-PMH service using Linked Data: the case of the Sheet Music Consortium [Electronic resource] / Stephen Davison // SWIB'13: Semantic Web in libraries, Nov. 25 – 27, 2013, Hamburg, Germany. — Electronic data. — 27 Nov. 2013. — Abstr., slides [29 p], video. — Mode of access: <http://swib.org/swib13/programme.php>, free. — Title of screen.
80. Cutter, C. A. Rules for a dictionary catalog / Charles A. Cutter, William P. Cutter, Worthington Ch. Ford [et al.] — 4th ed. — Washington : Gov. print. off., 1904. — 173 p. — (Special report on public libraries ; pt. 2).
81. Cyganiak, R. RDF 1.1: concepts and abstract syntax : W3C Recommendation 25 Febr. 2014 [Electronic resource] / 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.
82. Davis, I. RDF 1.1 JSON serialisation (RDF/JSON) : W3C Editor's Draft 07 Nov. 2013 [Electronic resource] / Ian Davis, Thomas Steiner, Arnaud J. le Hors // W3C: World Wide Web Consortium. — Electronic data. — Keio, Korea, Jap. ; Cambridge, MA, USA ; Biot, France ; Beihang, China, 2011—2013. — Mode of access: <https://dvcs.w3.org/hg/rdf/raw-file/default/rdf-json/index.html>, free. — Title of screen.
83. Debure, G.-F. Bibliographie instructive ou: traité de la connoissance des livres rares et singuliers / par Guillaume François de Bure, Louis Jean Gaignat, J-F Née de La Rochelle. — Paris : Chez Guillaume-François De Bure le jeune, 1763–1768. — 7 vol.
84. Dublin Core Metadata Initiative (DCMI): metadata innovation : [offic. website] [Electronic resource] / ASIS&T. — Electronic data. — [Silver Spring, Maryland, USA], [1995–2016]. — Mode of access: <http://dublincore.org/>, free. — Title of screen.
85. Dziatzko, K. Instruction für die Ordnung der Titel im alphabetischen Zettelkatalog der Königlicher und Universitäts-Bibliothek zu Breslau / Karl Dziatzko. — Berlin : Asher, 1886. — XI, 74 S.
86. Fellegi, I. P. A Theory for record linkage / Ivan P. Fellegi, Alan B. Sunter // J. of the Amer. Statist. Assoc. — 1969. — Vol. 64, N 328. — P. 1183—1210.

87. Flanagan, D. The Ruby programming language / David Flanagan, Yukihiro Matsumoto. — Beijing ; Sebastopol, CA : O'Reilly, 2008. — XI, 429 p. : ill.
88. Gradmann, S. Knowledge = Information in context: on the importance of semantic contextualisation in Europeana [Electronic resource] / Stefan Gradmann. — Electronic data. — [The Hague, Netherlands] : Europeana Office, 2010. — 19 p. : color ill., text file, PNG and PDF. — (Europeana white paper, 1 ; «April 2010»). — Mode of access: <http://webdoc.sub.gwdg.de/ebook/mon/2010/ppn%20630062188.pdf>, free. — Title of screen.
89. Guenther, R. S. MODS: the metadata object description schema / Rebecca S. Guenther // portal: Libr. a. the Acad. — 2003. — Vol. 3, N 1. — P. 137—150.
90. Halpin, H. When owl:sameAs isn't the same: an analysis of identity in Linked Data / Harry Halpin, Patrick J. Hayes, James P. McCusker [et al.] // The Semantic Web – ISWC'2010 : 9th intern. Semantic Web conf, Shanghai, China, Nov. 7 – 11, 2010 : rev. sel. papers. Pt. 1 / ed. by: Peter F. Patel-Schneider, Yue Pan, Pascal Hitzler [et al.]. — Berlin ; Heidelberg : Springer-Verlag, 2010. — P. 305—320. — (Lecture notes in computer science ; vol. 6496).
91. Halpin, H. Representing vCard Objects in RDF : W3C Member Submission 20 Jan. 2010 [Electronic resource] / Harry Halpin, Renato Iannella, Brian Suda, Norman Walsh ; ed. by Renato Iannella // W3C: World Wide Web Consortium. — Electronic data. — [Sydney, Australia], 2010. — Mode of access: <http://www.w3.org/Submission/vcard-rdf/>, free. — Title of screen.
92. Haslhofer, B. data.europeana.eu: the Europeana linked open data pilot [Electronic resource] / Bernhard Haslhofer, Antoine Isaac // DC-2011 – the Hague proceedings : intern. conf. on Dublin Core and Metadata Applications, the Hague, Netherlands, Sept. 21 – 23, 2011. — Electronic data. — [S.l.], 2011. — P. 94—104. — Mode of access: <http://dcpapers.dublincore.org/pubs/article/view/3625/1851>, free. — Title of screen.
93. Heath, T. Linked Data: evolving the Web into a global data space / Tom Heath, Christian Bizer. — [San Rafael, Calif.] : Morgan & Claypool, 2011. — xi, 122 p. : ill. — (Synthesis lectures on the Semantic Web: theory and technology ; vol. 1).

94. Hernández, M. A. The merge/purge problem for large databases / Maurisio A. Hernández, Salvatore J. Stolfo // ACM SIGMOD rec. — 1995. — Vol. 24, N 2. — P. 127—138.
95. Hickey, T. B. Description of the VIAF (Virtual International Authority File) dataset [Electronic resource] / Thomas B. Hickey, Jeffrey A. Young // SWJ: Semantic Web journal : spec. call for Linked Dataset description. — Electronic data. — 25.05.2012 — 6 p. — Mode of access: <http://www.semantic-web-journal.net/sites/default/files/swj294.pdf>, free. — Title of screen.
96. Hogan, A. Skolemising blank nodes while preserving isomorphism [Electronic resource] / 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.
97. 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.
98. Hylton, J. Identifying and merging related bibliographic records : [thesis/diss.] / Jeremy A. Hylton ; Massachusetts Inst. of Technology, Dept. of Electric. Engineering a. Computer Science. — Cambridge, MA, USA : Massachusetts Inst. of Technology, 1996. — 99 p. : ill.
99. Indyk, P. Approximate nearest neighbors: towards removing the curse of dimensionality / Piotr Indyk, Rajeev Motwani // Proceedings of the thirtieth annual ACM symposium on Theory of computing STOC'98, Dallas, TX, USA, May 24 – 26, 1998. — New York : ACM 1998. — P. 604—613.
100. Isaac, A. SKOS: simple knowledge organisation system / Antonie Isaac // IP: Inform. Professional. — 2006. — Vol. 10. — P. 40—43.
101. Isaac, A. Europeana linked open data – data.europeana.eu [Электронный ресурс] / Antonie Isaac, Bernhard Haslhofer // SWJ: Semantic Web journal : spec. call for

Linked Dataset description. — Electronic data. — 08.12.2012. — 7 p. — Mode of access: http://www.semantic-web-journal.net/system/files/swj297_1.pdf, free. — Title of screen.

102. Jaccard, P. Distribution de la flore alpine: dans le Bassin des Dranses et dans quelques régions voisines / Paul Jaccard // Bull. de la Soc. Vaudoise des sciences naturelles. — 1901. — Vol. XXXVII, N 140. — P. 241—272.

103. Jewett, C. C. Smithsonian report on the construction of catalogues of libraries, and of a general catalogue, and their publication by means of separate, stereotyped titles, with rules and examples / Charles Coffin Jewett. — Washington : Smithsonian institution, 1852. — 78 p. : octavo.

104. Johnson, G. Google Translate <http://translate.google.com/> / Gregory Johnson // Techn. services quart. — 2012. — Vol. 29, N 2. — P. 165.

105. Jurczyk, P. FRIL: a tool for comparative record linkage / Pavel Jurczyk, James J. Le, Li Xiong [et al.] // AMIA annu. symp. proc. / Amer. med. informatics assoc. — 2008. — Vol. 2008. — P. 440—444.

106. Khadilkar, V. Jena-HBase: a distributed, scalable and efficient RDF triple store / Vaibhav Khadilkar, Murat Kantarcioglu, Bhavani Thuraisingham, Paolo Castagna // CEUR workshop proc. — 2012. — Vol. 914. — P. 85—88.

107. 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.

108. Kramer, S. N. The oldest literary catalogue: a Sumerian List of literary compositions compiled about 2000 B.C. / Samuel N. Kramer // Bull. of the Amer. Schools of Oriental Research, 1942. — N 88. — P. 10—19.

109. Lewin, B. SOCH – the first cultural heritage Semantic Web : version 1.0 [Electronic resource] / Börje Lewin // SOCH report: project rep. from the course «Certifierad IT-arkitekt» N 44 at Dataföreningen Kompetens in Sweden — Electronic

data. — 2012/12/13. — 40 p. — Mode of access: http://www.ksamsok.se/wp-content/uploads/2013/04/Ksamsok_tek_spec1.pdf, free. — Title of screen.

110. A Catalogue of the Library of the College of St. Margaret and St. Bernard, commonly called Queen's College, in the University of Cambridge, methodically arranged by Thomas Hartwell Horne. Vol. II / Thomas Hartwell Horne ; Queen's College (University of Cambridge). — London : print. for the Soc. of Queens College ; S. a. R. Bentley, M.DCCC.XX.VII [1827]. — P. 514—1194.

111. Abrahams, P. V. TEX for the impatient / Paul V. Abrahams, Karl Berry, Kathryn Hargreaves. — Reading, Mass. : Addison-Wesley, 1992. —XVII, 357 p. : ill.

112. 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.

113. 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.

114. Mendes, P. N. DBpedia : a multilingual cross-domain knowledge base / Pablo N. Mendes, Max Jakob, Christian Bizer // LREC'2012 : proc. eighth intern. conf. on language resources and evaluation, Istanbul, [Turkey], May 21 – 27, 2012 / Europ. Language Resources Assoc. (ELRA). — [Paris], 2012. — P. 1813—1817.

115. Naudé G. Advis pour dresser une bibliothèque, présenté à monfigneur le président de Mesme. T. I / par Gabriel Naudé. — Paris : F. Targa, M.DC.XXVII, [1627]. — 167 p.

116. Ochiai, A. Zoogeographical studies on the soleoid fishes found in Japan and its neighbouring regions / Atsushi Ochiai // Bull. of the Jap. Soc. of Fish Sciences. — 1957. — Vol. 22, N 9. — P. 526—530.

117. Panizzi, A. G. M. Rules for the compilation of the catalogue / Antonio Genesio Maria Panizzi // Catalogue of printed books in the British Museum. Vol. 1. — London : J. B. Nichols a. Son, M.DCCC.XLI [1841]. — P. V-IX.
118. Pi, B. SimHash-based effective and efficient detecting of near-duplicate short messages / Bingfeng Pi, Shunkai Fu, Weilei Wang, Song Han // Proceedings the 2009 [second] International Symposium Computer Science and Computational Technology(ISCST'09), Huangshan, China, Dec. 26 – 28, 2009 / [ed. by: Fei Yu et al.]. — Oulu, Finland : [Acad. publ.], 2009. — P. 20—25.
119. Porter, M. F. An algorithm for suffix stripping / Martin F. Porter // Readings in information retrieval. — San Francisco, CA : Morgan Kaufmann publ. inc., 1997. — P. 313—316.
120. Powell, A. Metadata: collection level description. Collection description working group: work in progress : 12 Oct. 1998 [Electronic resource] / ed. by Andy Powell // UKOLN Metadata Publications. — Electronic data. — [Bath, UK, 1998]. — Mode of access: <http://www.ukoln.ac.uk/metadata/cld/wg-report/>, free. — Title of screen.
121. Reenskaug, T. The Model-View-Controller (MVC): its past and present [Electronic resource] / Trugve Reenskaug // University of Oslo, Department of Informatics : [webserver for all the personal homepages]. — Electronic data. — [Oslo, Norway], 2003. — 16 p. — Mode of access: http://home.ifi.uio.no/trygver/2003/javazone-jaoo/MVC_pattern.pdf, free. — Title of screen.
122. Shrivastava, A. In defense of MinHash over SimHash / Anshumali Shrivastava, Ping Li // The JMLR: Workshop and Conference Proceedings. Vol. 33 : Proceedings of the seventeenth international conference on artificial intelligence and statistics (AISTATS), Reykjavik, Iceland, Apr. 2, 2014 / [ed. by: Samuel Kaski, Jukka Corander]. — Cambridge, Mass. : [MIT Press], 2014. — P. 886—894.
123. Simon, A. Publishing bibliographic records on the Web of data: opportunities for the BnF (French National Library) / Agnès Simon, Romain Wenz, Vincent Michel, Adrien Di Mascio // The Semantic Web: semantics and big data : 10th intern. conf., ESWC'2013, Montpellier, France, May 26 – 30, 2013 : proc. / [ed. by: Philipp Cimiano,

Oscar Corcho, Valentina Presutti]. — Berlin ; Heidelberg : [Springer-Verlag], 2013. — P. 563—577. — (Lecture notes in computer science ; vol. 7882).

124. Svensson, L. G. National libraries and the Semantic Web: requirements and applications [Electronic resource] / Lars G. Svensson // CiteSeerX database / College of Inform. Sciences and Technology. — Electronic data. — [Pennsylvania State Univ., 2007]. — 8 p. — Mode of access: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.113.2756&rep=rep1&type=pdf>, free. — Title of screen.

125. Svensson, L. G. Practical Examples of the Use of CC0 or Linked Open Data from National Libraries by Third Parties : Examples from the Deutsche Nationalbibliothek (German National Library , DNB) / Lars G. Svensson // CENL Annual Meeting, Vaduz, Liechtenstein, Oct. 1, 2012. — 2012. — P 9—12.

126. Svensson, L. G. Are current bibliographic models suitable for integration with the Web? / Lars G. Svensson // Inform. standards quart. — 2013. — Vol. 25. — N 4. — P 6—13.

127. Definition of the Europeana Data Model, version 5.2.6, 17.12.2014 [Electronic resource] // Europeana.eu. Professional : [offic. website] / Europeana Foundation, Europ. Union. — Electronic data. — Den Haag, Netherlands, [2008—2016]. — 71 p. — Mode of access: http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM%20Definition%20v5.2.6_01032015.pdf, free. — Title of screen.

128. D5.3.1 – Europeana OAI-PMH infrastructure – documentation and final prototype Appendix – REPOX User Manual [Vienna, Austria, 11 Oct. 2010] [Electronic resource] / Gilberto Pedrosa, Petz Georg, Cesare Concordia, Nicola Aloia ; Europ. Union, Austr. Nat. Libr. // Europeana.eu. Connect : [offic. website] / Europeana Foundation, Europ. Union.. — Electronic data. — Den Haag, Netherlands, [2008—2010]. — 33 p. — Mode of access: http://www.europeanaconnect.eu/documents/01a_Europeana_OAI_PMH_APPENDIX_User%20Manual.pdf, free. — Title of screen.

129. Tillet, B. B. What is FRBR? A conceptual model for the bibliographic universe: [text originally published in Technicalities v. 25, no. 5 (Sept./Oct. 2003) Page [8]] / Barbara B. Tillet. — [Washington, D.C.] : Libr. of Congr., Cataloging Distribution Service, [2003]. — 7, [1] p. : ill.
130. Tillet, B. Keeping libraries relevant in the Semantic Web: RDA, VIAF, and the new bibliographic framework [Electronic resource] / Barbara Tillet ; Joint Steering Comm. for Development of RDA and Former Chief, Policy & Standards Division, Libr. of Congress // RDA Toolkit: Resource Description and Access / [Amer. Libr. Assoc., Canad. Libr. Assoc., CILIP: Chartered Inst. of Libr. a. Inform. Professionals]. — Electronic data. — Istanbul, Turkey, 19 Apr. 2013. — 41 p. — Mode of access: http://konferans.rdaturkiye.org/Barbara_Tillett.pdf, free. — Title of screen.
131. Trithemius, J. Liber de scriptoribus ecclesiasticis / Johannes Trithemius ; ed. Johannes de Lapide. — Basileae : J. Amerbach, 1494. — [6], 140, [1] fol.
132. Herman, I. RDFa 1.1 Primer Third Edition: W3C Working Group Note 17 March 2015 [Electronic resource] / Ivan Herman, Ben Adida, Manu Sporny, Mark Birbeck // W3C: World Wide Web Consortium. — Electronic data. — Keio, Korea, Jap. ; Cambridge, MA, USA ; Sophia Antipolis, France ; Beihang, China, 2003—2014. — Mode of access: <https://www.w3.org/TR/rdfa-primer/>, free. — Title of screen.
133. Winkler, W. E. Overview of record linkage and current research directions / William E. Winkler ; Stat. Research Div., U.S. Census Bureau. — Washington, 2006. — 44 p.
134. Xiao, C. Efficient similarity joins for near duplicate detection / Chuan Xiao, Wei Wang, Xuemin Lin, Jeffrey Xu Yu // Proceedings of the 17th international conference on World Wide Web, Beijing, China, April 21 – 25, 2008. — New York : ACM 2008. — P. 131—140
135. Young, J. OAICat [Electronic resource] / Jeff Young // OCLC Research. — Electronic data. — [Leiden, Netherlands]. — Page content: 2009-08-11. — Prototype: 2003-02-03. — Mode of access: <http://www.oclc.org/research/themes/data-science/oaicat.html>, free. — Title of screen.

136. Free Data Services [Electronic resource] // The British Library Collection Metadata. — Electronic data. — Boston, UK, 2012. — Mode of access: <http://www.bl.uk/bibliographic/datafree.html>, free. — Title of screen.
137. VIAF. Virtual International Authority File [Electronic resource] // OCLC. — Electronic data. — [Leiden, Netherlands], 2012. — Mode of access: <http://www.oclc.org/viaf.en.html>, free. — Title of screen.
138. GeoNames [Electronic resource] // GeoNames : [the geogr. database] / Unxos GmbH. — Electronic data. — [Zürich], Switzerland. — Mode of access: <http://www.geonames.org/about.html>, free. — Title of screen.
139. About GEMET – GEneral Multilingual Environmental [Electronic resource] / Europ. Environment Agency // Eionet: GEMET Thesaurus. — Electronic data. — [Copenhagen, Norway], Mar. 12, 2008, last version: Oct. 10, 2013. — Mode of access: <https://www.eionet.europa.eu/gemet/about?langcode=fr>, free. — Title of screen.
140. DC-Library Application Profile (DC-Lib) [Электронный ресурс] / [Robina Clauphan, Rebekka Guenther] ; DCMI-Libr. Working Group // Dublin Core Metadata Initiative (DCMI): metadata innovation : [offic. website] / ASIS&T. — Electronic data. — [Silver Spring, Maryland, USA]. — 2004/09/10. — Mode of access: <http://dublincore.org/documents/library-application-profile/>, free. — Title of screen.
141. Новости Министерства. Художественное училище памяти 1905 года обновило материально-техническую базу впервые за 30 лет : [итоги заседания обновл. коллегии М-ва культуры Рос. Федерации, 24.04.2014] [Электронный ресурс] // Министерство Культуры Российской Федерации : офиц. сайт. — Электрон. дан. — [Москва], 2014. — Режим доступа: <http://mkrf.ru/m/494838/>, свободный. — Загл. с экрана.
142. Концепция Сводного каталога электронных ресурсов (СКЭР) [Электронный ресурс] // Национальный информационно-библиотечный центр ЛИБНЕТ : [офиц. портал]. — Электрон. дан. — [Москва, 2012]. — 2 с. — Режим доступа: http://www.nilc.ru/nilc/documents/konsepsia_sker.pdf, свободный. — Загл. с экрана.

143. AIDA64 [Electronic resource] / FinalWire Ltd. — Electronic data. — [Budapest, Hungary], 2010—2016. — Mode of access: <http://www.aida64.com>, free. — Title of screen.
144. Российский коммуникативный формат представления библиографических записей в машиночитаемой форме: (русская версия UNIMARC) : [утв. приказом М-ва культуры Рос. Федерации № 45 от 27.01.1998 г.] [Электронный ресурс] / М-во культуры Рос. Федерации, Рос. Библ. Ассоц. // Российский коммуникативный формат RUSMARC: [офиц. сайт]. — Электрон. дан. — Москва ; Санкт-Петербург. — 22.09.2011. — Режим доступа: <http://www.rusmarc.ru/rusmarc/format.html>, свободный. — Загл. с экрана.
145. Speedtest: [the global broadband speed test for Internet connection] [Electronic resource] / Ookla: the global standard in internet metrics. — Electronic data. — [Seattle, WA.], 2016. — Mode of access: <http://www.speedtest.net>, free. — Title of screen.
146. 4store: [an RDF database] [Electronic resource] / 4store: an RDF database. — Electronic data. — [Costa Mesa, CA, USA], 2016. — Mode of access: <http://4store.org>, free. — Title of screen.
147. Sesame : [an open source Java framework for processing and handling RDF data] [Electronic resource] / Callimachus. — Electronic data. — [Arlington, VA, USA], 2016. — Mode of access: <http://rdf4j.org>, free. — Title of screen.
148. Apache Jena: a free and open source Java framework for building Semantic Web and Linked Data applications [Electronic resource] / Apache Software Foundation. — Electronic data. — [Forest Hill, MD, USA], 2011—2015. — Mode of access: <https://jena.apache.org>, free. — Title of screen.
149. DB-Engines Ranking [Electronic resource] // DB-Engines: knowledge base of relational and NoSQL database management systems / solid IT GmbH. — Electronic data. — [Heilbronn, Germany], 2016. — Mode of access: <http://db-engines.com/en/ranking>, free. — Title of screen.
150. Apache Jena Fuseki: Fuseki2 is a new major revision of Fuseki [Electronic resource] // Apache Jena: a free and open source Java framework for building Semantic

- Web and Linked Data applications / Apache Software Foundation. — Electronic data. — [Forest Hill, MD, USA], 2011—2015. — Mode of access: <http://jena.apache.org/documentation/fuseki2/index.html>, free. — Title of screen.
151. SPARQL 1.1 Protocol : W3C Recommendation 21 Mar. 2013 [Electronic resource] / ed. by: Lee Feigenbaum, Gregory Todd Williams, Kendall Grant Clark, Elias Torres // W3C: World Wide Web Consortium. — Electronic data. — Keio, Korea, Jap. ; Cambridge, MA, USA ; Sophia Antipolis, France ; Beihang, China, 2013. — Mode of access: <http://www.w3.org/TR/sparql11-protocol/>, free. — Title of screen.
152. Play 2.4.x documentation [Electronic resource] // Play Framework: [the high velocity Web Framework for Java and Scala] / Singularity SA. — Electronic data. — [Paris, France, 2016]. — Mode of access: <https://www.playframework.com/documentation/2.4.x/Home>, free. — Title of screen.
153. Numerus librorum Glastoniensis ecclesiae, qui fuerunt de libraria anno 1247 // Chronica sive historia de rebus glastoniensibus. [2 vol. in 1] / Johannes Glastoniensis ; [ed. Th. Hearnii]. — Oxonii : [e theatro Sheldoniano], 1726. — Vol. II. — P. 423—444.
154. PRISM: Publishing Requirements for Industry Standard Metadata [Electronic resource] // Idealliance (International Digital Enterprise Alliance). — Electronic data. — Alexandria, VA, [USA], 2004—2015. — Mode of access: <http://www.idealliance.org/specifications/prism-metadata-initiative/prism>, free. — Title of screen.
155. Europeana.eu. Collections : [offic. website] [Electronic resource] / Europeana Foundations, Europ. Union. — Electronic data. — Den Haag, Netherlands, [2008—2016]. — Mode of access: <http://www.europeana.eu/portal/>, free. — Title of screen.
156. Facts & Figures [Electronic resource] // Europeana inside. — Electronic data. — [London], 2012. — <http://www.europeana-inside.eu/home/index.html>, free. — Title of screen.

Приложения

Приложение А

Описание элементов словаря Europeana Semantic Elements

alternative	
Пространство имен	dcterms
URI	http://purl.org/dc/terms/alternative
Название	Альтернативное название
Описание	Альтернативное имя для ресурса. Это может быть любая форма названия, которая может быть использована в качестве замены или альтернативы формальному названию ресурса, включая аббревиатуры, перевод и т.п. Является уточнением dc:title.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

conformsTo	
Пространство имен	dcterms
URI	http://purl.org/dc/terms/conformsTo
Название	Соответствует
Описание	Устанавливает стандарт, которому соответствует описываемый ресурс. Является уточнением dc:relation.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

contributor	
Пространство имен	dc
URI	http://purl.org/dc/elements/1.1/contributor
Название	Поставщик

Описание	Персона, организация или сервис, отвечающий за наполнение ресурса.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

country	
Пространство имен	europeana
URI	http://www.europeana.eu/schemas/ese/country
Название	Страна
Описание	Название страны поставщика в формате ISO 3166. Для проектов, охватывающих несколько стран – Europe.
Ограничения	Обязательный элемент. Минимальное количество – 1, максимальное – 1.

coverage	
Пространство имен	dc
URI	http://purl.org/dc/elements/1.1/coverage
Название	Охват
Описание	Пространственный или временной аспект описания ресурса, пространственная применимость ресурса, или юрисдикция, где ресурс актуален. Это может быть географическое название, пространственные координаты, период, дата, временной отрезок. Уточняется с помощью dcterms:spatial и dcterms:temporal.
Ограничения	Обязательно должен быть указан dc:coverage или dcterms:spatial или dc:subject или dc:type. Минимальное количество – 0, максимальное – без ограничений.

created

Пространство имен	dcterms
URI	http://purl.org/dc/terms/created
Название	Дата создания
Описание	Дата создания ресурса. Является уточнением dc:date.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

creator	
Пространство имен	dc
URI	http://purl.org/dc/elements/1.1/creator
Название	Создатель
Описание	Персона, организация или сервис, отвечающий за создание ресурса.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

dataProvider	
Пространство имен	europeana
URI	http://www.europeana.eu/schemas/ese/dataProvider
Название	Поставщик данных в Europeana
Описание	Имя или идентификатор организации, поставляющий данные в Europeana.
Ограничения	Обязательный элемент. Минимальное количество – 1, максимальное – 1.

date	
Пространство имен	dc
URI	http://purl.org/dc/elements/1.1/date
Название	Дата

Описание	Дата или период времени, ассоциированный с событием в жизненном цикле ресурса. Уточняется с помощью dcterms:created и dcterms:issued.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

description	
Пространство имен	dc
URI	http://purl.org/dc/elements/1.1/description
Название	Описание
Описание	Текстовое описание ресурса. Уточняется с помощью dcterms:tableOfContents.
Ограничения	Обязательно должен быть указан dc:description или dc:title. Минимальное количество – 0, максимальное – без ограничений.

extent	
Пространство имен	dcterms
URI	http://purl.org/dc/terms/extent
Название	Размер
Описание	Размер ресурса. Является уточнением dc:format.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

format	
Пространство имен	dc
URI	http://purl.org/dc/elements/1.1/format
Название	Формат
Описание	Формат файла, физический носитель или размер ресурса.

	Уточняется с помощью dcterms:extent и dcterms:medium.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

hasFormat	
Пространство имен	dcterms
URI	http://purl.org/dc/terms/hasFormat
Название	Имеет формат
Описание	Взаимосвязанный ресурс, который в значительной степени является тем же, что и существующий описываемый ресурс, но в другом формате. Является уточнением dc:relation.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

hasPart	
Пространство имен	dcterms
URI	http://purl.org/dc/terms/hasPart
Название	Имеет часть
Описание	Взаимосвязанный ресурс, который логически или физически входит в состав описываемого ресурса. Является уточнением dc:relation.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

hasVersion	
Пространство имен	dcterms
URI	http://purl.org/dc/terms/hasVersion
Название	Имеет версию
Описание	Взаимосвязанный ресурс, который является версией,

	изданием или адаптацией описываемого ресурса. Изменение версии подразумевает изменение контента, а не формата ресурса. Является уточнением dc:relation.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

identifier	
Пространство имен	dc
URI	http://purl.org/dc/elements/1.1/identifier
Название	Идентификатор
Описание	Однозначный указатель на ресурс.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

isFormatOf	
Пространство имен	dcterms
URI	http://purl.org/dc/terms/isFormatOf
Название	Является форматом
Описание	Взаимосвязанный ресурс, который в значительной степени является тем же, что и описываемый ресурс, но в другом формате. Является уточнением dc:relation.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

isPartOf	
Пространство имен	dcterms
URI	http://purl.org/dc/terms/isPartOf
Название	Является частью
Описание	Взаимосвязанный ресурс, в состав которого логически или

	физически входит описываемый ресурс. Является уточнением dc:relation.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

isReferencedOf	
Пространство имен	dcterms
URI	http://purl.org/dc/terms/isReferencedBy
Название	Указывается
Описание	Взаимосвязанный ресурс, который указывает, цитирует или каким-либо образом ссылается на описываемый ресурс. Является уточнением dc:relation.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

isReplacedOf	
Пространство имен	dcterms
URI	http://purl.org/dc/terms/isReplacedBy
Название	Замещается
Описание	Взаимосвязанный ресурс, который вытесняет, замещает или отменяет собой описываемый ресурс. Является уточнением dc:relation.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

isRequiredOf	
Пространство имен	dcterms
URI	http://purl.org/dc/terms/isRequiredBy
Название	Требуется

Описание	Взаимосвязанный ресурс, для которого требуется, чтобы описываемый ресурс поддерживал функционирование. Является уточнением dc:relation.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

isShownAt	
Пространство имен	europeana
URI	http://www.europeana.eu/schemas/ese/isShownAt
Название	Показан в
Описание	Однозначная URL ссылка на цифровой объект на веб-сайт провайдера, где можно получить максимально возможную информацию об объекте. Является уточнением dc:relation.
Ограничения	Обязательным является использование europeana:isShownAt или europeana:isShownBy. Минимальное количество – 0, максимальное – 1.

isShownBy	
Пространство имен	europeana
URI	http://www.europeana.eu/schemas/ese/isShownBy
Название	Показан
Описание	Однозначная URL ссылка на цифровой объект на веб-сайт провайдера, где ресурс представлен в максимально возможном качестве/разрешении. Является уточнением dc:relation.
Ограничения	Обязательным является использование europeana:isShownAt или europeana:isShownBy. Минимальное количество – 0, максимальное – 1.

issued	
Пространство имен	dcterms
URI	http://purl.org/dc/terms/issued
Название	Дата выпуска
Описание	Дата формального выпуска (публикации) ресурса. Является уточнением dc:date.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

isVersionOf	
Пространство имен	dcterms
URI	http://purl.org/dc/terms/isVersionOf
Название	Является версией
Описание	Взаимосвязанный ресурс, для которого описываемый ресурс является версией, изданием или адаптацией. Изменение версии подразумевает изменение контента, а не формата ресурса. Является уточнением dc:relation.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

language	
Пространство имен	dc
URI	http://purl.org/dc/elements/1.1/language
Название	Язык
Описание	Язык ресурса. Рекомендуется использовать контролируемый словарь двух- и трех-буквенных обозначений.
Ограничения	Является обязательным элементом для текстовых объектов. Минимальное количество – 0, максимальное – без ограничений.

language	
Пространство имен	europaana
URI	http://www.europeana.eu/schemas/ese/language
Название	Язык в Europaana
Описание	Стандартизованное представление имени языка, которое используется для фасетного поиска на портале.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

medium	
Пространство имен	dcterms
URI	http://purl.org/dc/terms/medium
Название	Носитель
Описание	Материал или физический носитель ресурса. Является уточнением dc:format.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

object	
Пространство имен	europaana
URI	http://www.europeana.eu/schemas/ese/object
Название	Объект
Описание	Ссылка на изображение в максимальном качестве/разрешении, из которого могут быть автоматически сгенерированы изображения уменьшенных размеров для нужд портала.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – 1.

provenance	
Пространство имен	dcterms
URI	http://purl.org/dc/terms/provenance
Название	Происхождение
Описание	Содержит информацию обо всех изменениях собственности и владения ресурсом с момента его создания.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

provider	
Пространство имен	europeana
URI	http://www.europeana.eu/schemas/ese/provider
Название	Поставщик в Europeana
Описание	Название организации, поставляющей данные в Europeana.
Ограничения	Является обязательным элементом. Минимальное количество – 1, максимальное – 1.

publisher	
Пространство имен	dc
URI	http://purl.org/dc/elements/1.1/publisher
Название	Издатель
Описание	Персона, организация или сервис, ответственный за выпуск ресурса.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

references	
Пространство имен	dcterms

URI	http://purl.org/dc/terms/references
Название	Указывает
Описание	Взаимосвязанный ресурс, на который указывает, цитирует или каким-либо образом ссылается описываемый ресурс. Является уточнением dc:relation.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

relation	
Пространство имен	dc
URI	http://purl.org/dc/elements/1.1/relation
Название	Взаимосвязь
Описание	Взаимосвязанный ресурс. Уточняется с использованием dcterms:isVersionOf, dcterms:hasVersion, dcterms:isReplacedBy, dcterms:replaces, dcterms:isRequiredBy, dcterms:requires, dcterms:isPartOf, dcterms:hasPart, dcterms:isReferencedBy, dcterms:references, dcterms:isFormatOf, dcterms:hasFormat, dcterms:conformsTo, europeana:isShownBy, и europeana:isShownAt.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

replaces	
Пространство имен	dcterms
URI	http://purl.org/dc/terms/replaces
Название	Замещает
Описание	Взаимосвязанный ресурс, который вытесняется, замещается или отменяется описываемым ресурсом. Является уточнением dc:relation.

Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.
-------------	---

requires	
Пространство имен	dcterms
URI	http://purl.org/dc/terms/requires
Название	Требует
Описание	Взаимосвязанный ресурс, который требуется для того, чтобы описываемый ресурс поддерживал функционирование. Является уточнением dc:relation.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

rights	
Пространство имен	dc
URI	http://purl.org/dc/elements/1.1/rights
Название	Права
Описание	Информация о правах на ресурс.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

rights	
Пространство имен	europeana
URI	http://www.europeana.eu/schemas/ese/rights
Название	Права в Europeana
Описание	Информация о правах на цифровой объект, указанный в isShownBy и isShownAt, а также для изображений, используемых для предпросмотра на портале.
Ограничения	Является обязательным элементом. Минимальное

	количество – 1, максимальное – 1.
--	-----------------------------------

source	
Пространство имен	dc
URI	http://purl.org/dc/elements/1.1/source
Название	Источник
Описание	Взаимосвязанный ресурс, из которого происходит описываемый ресурс.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

spatial	
Пространство имен	dcterms
URI	http://purl.org/dc/terms/spatial
Название	Пространственный охват
Описание	Пространственные характеристики ресурса. Является уточнением dc:coverage.
Ограничения	Обязательно должен быть указан dc:coverage или dcterms:spatial или dc:subject или dc:type. Минимальное количество – 0, максимальное – без ограничений.

subject	
Пространство имен	dc
URI	http://purl.org/dc/elements/1.1/subject
Название	Предмет
Описание	Тема ресурса.
Ограничения	Обязательно должен быть указан dc:coverage или dcterms:spatial или dc:subject или dc:type. Минимальное количество – 0, максимальное – без ограничений.

tableOfContents	
Пространство имен	dcterms
URI	http://purl.org/dc/terms/tableOfContents
Название	Таблица содержимого
Описание	Список элементов, из которых состоит ресурс. Является уточнением dc:description.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

temporal	
Пространство имен	dcterms
URI	http://purl.org/dc/terms/temporal
Название	Временной охват
Описание	Временные характеристики ресурса. Является уточнением dc:coverage.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

title	
Пространство имен	dc
URI	http://purl.org/dc/elements/1.1/title
Название	Заголовок
Описание	Имя, данное ресурсу. Уточняется с помощью dcterms:alternative.
Ограничения	Обязательно должен быть указан dc:description или dc:title. Минимальное количество – 0, максимальное – без ограничений.

type	
Пространство имен	dc
URI	http://purl.org/dc/elements/1.1/type
Название	Тип
Описание	Тип ресурса, который обычно включает в себя информацию о жанре, категориях, уровнях агрегации и функциях контента.
Ограничения	Обязательно должен быть указан dc:coverage или dcterms:spatial или dc:subject или dc:type. Минимальное количество – 0, максимальное – без ограничений.

type	
Пространство имен	europeana
URI	http://www.europeana.eu/schemas/ese/type
Название	Тип в Europeana
Описание	Тип ресурса в Europeana.
Ограничения	Является обязательным элементом. Минимальное количество – 1, максимальное – 1.

ugc	
Пространство имен	europeana
URI	http://www.europeana.eu/schemas/ese/ugc
Название	Контент, созданный пользователем
Описание	Этот элемент используется для идентификации контента, созданного пользователем. Этот элемент должен сопровождать все объекты, которые были собраны в процессе краудсорсинга.
Ограничения	Является обязательным элементом в том случае, если он применим к объекту. Минимальное количество – 0,

	максимальное – 1.
--	-------------------

unstored	
Пространство имен	europeana
URI	http://www.europeana.eu/schemas/ese/unstored
Название	Unstored
Описание	Контейнер, содержащий информацию, которую нельзя сохранить с использованием других конструкций ESE. Используется для хранения информации, которая может быть важна для создания поискового индекса. Не рекомендуется использовать.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

uri	
Пространство имен	europeana
URI	http://www.europeana.eu/schemas/ese/uri
Название	URI в Europeana
Описание	Однозначный идентификатор ресурса внутри Europeana.
Ограничения	Является обязательным элементом. Генерируется Europeana. Минимальное количество – 1, максимальное – 1.

userTag	
Пространство имен	europeana
URI	http://www.europeana.eu/schemas/ese/userTag
Название	Пользовательская метка
Описание	Метка, создаваемая пользователем с помощью интерфейса портала Europeana.
Ограничения	Необязательный элемент. Минимальное количество – 0,

	максимальное – без ограничений.
--	---------------------------------

year	
Пространство имен	europeana
URI	http://www.europeana.eu/schemas/ese/year
Название	Год
Описание	Четырехзначное обозначение года, когда произошло какое-либо событие, связанное с ресурсом.
Ограничения	Необязательный элемент. Минимальное количество – 0, максимальное – без ограничений.

Приложение Б

Свойства основных классов Europeana Data Model

Обозначения:

- **Полужирным** шрифтом обозначены свойства, которые являются обязательными.
- *Курсивом* выделены свойства, которые рекомендуется использовать.
- Одно из свойств, обозначенных с помощью (1), должно быть указано.
- Одно из свойств, обозначенных с помощью (2), должно быть указано.

Свойства класса edm:ProvidedCHO

dc:contributor	dcterms:created	dcterms:spatial (1)
dc:coverage (1)	dcterms:extent	dcterms:tableOfContents
dc:creator	dcterms:hasFormat	dcterms:temporal
dc:date	dcterms:hasPart	edm:currentLocation
dc:description (2)	dcterms:hasVersion	edm:hasMet
dc:format	dcterms:isFormatOf	edm:hasType
dc:identifier	dcterms:isPartOf	edm:incorporates
dc:language (если edm:type = TEXT)	dcterms:isReferencedBy	edm:isDerivativeOf
dc:publisher	dcterms:isReplacedBy	edm:isNextInSequence
dc:relation	dcterms:isRequiredBy	edm:isRelatedTo
dc:rights	dcterms:issued	edm:isRepresentationOf
dc:source	dcterms:isVersionOf	edm:isSimilarTo
dc:subject (1)	dcterms:medium	edm:isSuccessorOf
dc:title (2)	dcterms:provenance	edm:realizes
dc:type (1)	dcterms:references	edm:type
dcterms:alternative	dcterms:replaces	owl:sameAs
dcterms:conformsTo	dcterms:requires	

Свойства класса edm:WebResource

dc:creator	dcterms:confirmsTo	dcterms:isPartOf
dc:description	dcterms:created	dcterms:issued
dc:format	dcterms:extent	edm:isNextInSequence
dc:rights	dcterms:hasPart	edm:rights
dc:source	dcterms:isFormatOf	owl:sameAs

Свойства класса ore:Aggregation

edm:aggregatedCHO	edm:isShownBy (1)	edm:rights
edm:dataProvider	edm:object	edm:ugc
edm:hasView	edm:provider	
edm:isShownAt (1)	dc:rights	

Свойства класса edm:Agent

skos:prefLabel	edm:end	rdaGr2:dateOfTermination
skos:altLabel	edm:hasMet	rdaGr2:gender
skos:note	edm:isRelatedTo	rdaGr2:placeOfBirth
dc:date	foaf:name	rdaGr2:placeOfDeath
dc:identifier	rdaGr2:biographicalInformation	rdaGr2:professionOrOccupation
dcterms:hasPartOf	rdaGr2:dateOfBirth	owl:sameAs
dcterms:isPartOf	rdaGr2:dateOfDeath	
edm:begin	rdaGr2:dateOfEstablishment	

Свойства класса edm:Place

wgs84_pos:lat	skos:altLabel	edm:isNextInSequence
wgs84_pos:long	skos:note	owl:sameAs
wgs84_pos:alt	dcterms:hasPartOf	
skos:prefLabel	dcterms:isPartOf	

Свойства класса edm:TimeSpan

skos:prefLabel	dcterms:hasPartOf	edm:end
skos:altLabel	dcterms:isPartOf	edm:isNextInSequence
skos:note	edm:begin	owl:sameAs

Свойства класса skos:Concept

skos:prefLabel	skos:broadMatch	skos:note
skos:altLabel	skos:narrowMatch	skos:notation
skos:broader	skos:relatedMatch	skos:inScheme
skos:narrower	skos:exactMatch	
skos:related	skos:closeMatch	

Свойства класса cc:License

odrl:inheritFrom	cc:deprecatedOn	
------------------	-----------------	--

Уточнение свойств классов Europeana Data Model

Свойство	Уточняющее свойство
dcterms:provenance	edm:provider; edm:dataProvider
edm:isRelatedTo	dc:relation; edm:realizes; edm:hasType; edm:begin; edm:end; dc:subject
dc:relation	edm:occuredAt; edm:happenedAt; edm:wasPresentAt; edm:isSimilarTo; edm:isNextInSequence; edm:hasMet; dcterms:hasPart
dc:subject	edm:isAnnotationOf; edm:aggregatedCHO; edm:isRepresentationOf
edm:isSimilarTo	edm:isDerivativeOf; edm:isSuccessorOf; edm:incorporates
edm:hasMet	edm:currentLocation
dcterms:hasPart	ore:aggregates
ore:aggregates	edm:hasView; edm:landingPage; edm:aggregatedCHO
edm:hasView	edm:isShownAt; edm:isShownBy; edm:preview

Приложение В

Пример библиографической записи в формате MARC21

Поле	Индикатор	Подполе	Значение
LDR			01241cam#a2200229#i#4500
001			003538908
005			20080702124604.0
008			080422s2008####ru#####000# #rus d
017	##	\$a	08-28354
020	##	\$a	978-5-486-02129-9 (В пер.)
040	##	\$a	RuMoRKP
		\$b	rus
		\$e	rcr
		\$d	RuMoRGB
041	0#	\$a	rus
080	##	\$a	821.161.1-31
084	##	\$a	III5(2=P)52-644
		\$2	rubbk
100	1#	\$a	Достоевский, Федор Михайлович
		\$d	1821-1881
245	00	\$a	Преступление и наказание :
		\$b	роман в шести частях с эпилогом
		\$c	Федор Михайлович Достоевский
260	##	\$a	Москва
		\$b	Мир книги
		\$b	Литература
		\$c	2008
300	##	\$a	510, [1] с.
		\$c	21 см
490	0#	\$a	Бриллиантовая коллекция

650	#7	\$a	Филологические науки. Художественная литература -- Россия -- Русская литература -- 2-ая пол. 19 в. (до сер. 90-х гг.) -- Произведения художественной литературы -- Художественная проза -- Романы. Повести. Рассказы
		\$2	rubbk
852	##	\$a	РГБ
		\$b	ORF
		\$x	82
852	##	\$a	РГБ
		\$b	FB
		\$j	10 08-20/305
		\$x	82

Пример библиографической записи в формате MARCXML

```

<collection xmlns="http://www.loc.gov/MARC21/slim">
  <record>
    <leader>01251cam a2200241 i 4500</leader>
    <controlfield tag="001">003538908</controlfield>
    <controlfield tag="005">20080702124604.0</controlfield>
    <controlfield tag="008">080422s2008 ru 000 | rus|d</controlfield>
    <datafield tag="017" ind1=" " ind2=" ">
      <subfield code="a">08-28354</subfield>
    </datafield>
    <datafield tag="020" ind1=" " ind2=" ">
      <subfield code="a">978-5-486-02129-9 (В пер.)</subfield>
    </datafield>
    <datafield tag="040" ind1=" " ind2=" ">

```

```

<subfield code="a">RuMoRKP</subfield>
<subfield code="b">rus</subfield>
<subfield code="e">rcr</subfield>
<subfield code="d">RuMoRGB</subfield>
</datafield>
<datafield tag="041" ind1="0" ind2=" ">
  <subfield code="a">rus</subfield>
</datafield>
<datafield tag="080" ind1=" " ind2=" ">
  <subfield code="a">821.161.1-31</subfield>
</datafield>
<datafield tag="084" ind1=" " ind2=" ">
  <subfield code="a">III5(2=P)52-644</subfield>
  <subfield code="2">rubbk</subfield>
</datafield>
<datafield tag="100" ind1="1" ind2=" ">
  <subfield code="a">Достоевский, Федор Михайлович</subfield>
  <subfield code="d">1821-1881</subfield>
</datafield>
<datafield tag="245" ind1="0" ind2="0">
  <subfield code="a">Преступление и наказание :</subfield>
  <subfield code="b">роман в шести частях с эпилогом</subfield>
  <subfield code="c">Федор Михайлович Достоевский</subfield>
</datafield>
<datafield tag="260" ind1=" " ind2=" ">
  <subfield code="a">Москва</subfield>
  <subfield code="b">Мир книги</subfield>
  <subfield code="b">Литература</subfield>
  <subfield code="c">2008</subfield>
</datafield>

```

<datafield tag="300" ind1=" " ind2=" ">

<subfield code="a">510, [1] с.</subfield>

<subfield code="c">21 см</subfield>

</datafield>

<datafield tag="490" ind1="0" ind2=" ">

<subfield code="a">Бриллиантовая коллекция</subfield>

</datafield>

<datafield tag="650" ind1=" " ind2="7">

<subfield code="a">

Филологические науки. Художественная литература -- Россия -- Русская литература -- 2-ая пол. 19 в. (до сер. 90-х гг.) -- Произведения художественной литературы -- Художественная проза -- Романы. Повести. Рассказы

</subfield>

<subfield code="2">rubbk</subfield>

</datafield>

<datafield tag="852" ind1=" " ind2=" ">

<subfield code="a">RGB</subfield>

<subfield code="b">FB</subfield>

<subfield code="j">10 08-20/305</subfield>

</datafield>

<datafield tag="852" ind1=" " ind2=" ">

<subfield code="a">RGB</subfield>

<subfield code="b">ORF</subfield>

</datafield>

<datafield tag="912" ind1=" " ind2=" ">

<subfield code="a">08-1570</subfield>

</datafield>

</record>

</collection>

Пример библиографической записи в формате MODS

```

<mods
          xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://www.loc.gov/mods/v3"
xsi:schemaLocation=http://www.loc.gov/mods/v3
http://www.loc.gov/standards/mods/v3/mods-3-5.xsd version="3.5">
  <titleInfo>
    <title>Преступление и наказание</title>
    <subTitle>роман в шести частях с эпилогом</subTitle>
  </titleInfo>
  <name type="personal" usage="primary">
    <namePart>Достоевский, Федор Михайлович</namePart>
    <namePart type="date">1821-1881</namePart>
  </name>
  <typeOfResource>text</typeOfResource>
  <originInfo>
    <place>
      <placeTerm type="code" authority="marccountry">ru</placeTerm>
    </place>
    <place>
      <placeTerm type="text">Москва</placeTerm>
    </place>
    <publisher>Мир книги</publisher>
    <publisher>Литература</publisher>
    <dateIssued>2008</dateIssued>
    <issuance>monographic</issuance>
  </originInfo>
  <language>
    <languageTerm authority="iso639-2b" type="code">rus</languageTerm>
  </language>

```

<physicalDescription>

<form authority="marcform">print</form>

<extent>510, [1] с. 21 см</extent>

</physicalDescription>

<note type="statement of responsibility" altRepGroup="00">Федор Михайлович Достоевский</note>

<subject authority="rubbk">

<topic>

Филологические науки. Художественная литература -- Россия -- Русская литература -- 2-ая пол. 19 в. (до сер. 90-х гг.) -- Произведения художественной литературы -- Художественная проза -- Романы. Повести. Рассказы

</topic>

</subject>

<classification authority="udc">821.161.1-31</classification>

<classification authority="rubbk">III5(2=P)52-644</classification>

<location>

<physicalLocation>RGB FB</physicalLocation>

<shelfLocator>10 08-20/305</shelfLocator>

</location>

<location>

<physicalLocation>RGB ORF</physicalLocation>

</location>

<relatedItem type="series">

<titleInfo>

<title>Бриллиантовая коллекция</title>

</titleInfo>

</relatedItem>

<identifier type="isbn">978-5-486-02129-9 (В пер.)</identifier>

<recordInfo>

<descriptionStandard>rcr</descriptionStandard>

```

<recordContentSource authority="marcorg">RuMoRKP</recordContentSource>
<recordCreationDate encoding="marc">080422</recordCreationDate>
<recordChangeDate encoding="iso8601">20080702124604.0</recordChangeDate>
<recordIdentifier>003538908</recordIdentifier>
<languageOfCataloging>
  <languageTerm authority="iso639-2b" type="code">rus</languageTerm>
</languageOfCataloging>
</recordInfo>
</mods>

```

Пример библиографической записи в формате MODS/RDF

```

<rdf:RDF
    xmlns:fo="http://www.w3.org/1999/XSL/Format"
    xmlns:xs="http://www.w3.org/2001/XMLSchema"
    xmlns:fn="http://www.w3.org/2005/xpath-functions"
    xmlns:mods="http://www.loc.gov/mods/v3"
    xmlns:modsrdf="http://www.loc.gov/mods/rdf/v1#"
    xmlns:madsrdf="http://www.loc.gov/mads/rdf/v1#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:identifier="http://id.loc.gov/vocabulary/identifier/"
    xmlns:relator="http://id.loc.gov/vocabulary/relator/"
    xmlns:note="http://id.loc.gov/vocabulary/note/"
    xmlns:abstract="http://id.loc.gov/vocabulary/abstract/"
    xmlns:access="http://id.loc.gov/vocabulary/access/"
    xmlns:class="http://id.loc.gov/vocabulary/class/"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:ri="http://id.loc.gov/ontologies/RecordInfo#"
    xmlns:xlink="http://www.w3.org/1999/xlink">
  <modsrdf:ModsResource
    rdf:about="http://www.loc.gov/mods/rdf/v1#MODS123456">

```

```

<rdf:type rdf:resource="http://id.loc.gov/vocabulary/resourceType#Text"/>
<namePrincipal xmlns="http://www.loc.gov/mods/rdf/v1#">
  <PersonalName xmlns="http://www.loc.gov/mads/rdf/v1#" rdf:about="d1">
    <label xmlns="http://www.w3.org/2000/01/rdf-schema#">Достоевский, Федор
Михайлович</label>
    <elementList rdf:parseType="Collection">
      <FullNameElement>
        <elementValue>Достоевский, Федор Михайлович 1821-
1881</elementValue>
      </FullNameElement>
      <DateNameElement>
        <elementValue>1821-1881</elementValue>
      </DateNameElement>
    </elementList>
  </PersonalName>
</namePrincipal>
<titlePrincipal xmlns="http://www.loc.gov/mods/rdf/v1#">
  <Title xmlns="http://www.loc.gov/mads/rdf/v1#">
    <label xmlns="http://www.w3.org/2000/01/rdf-schema#">
Преступление и наказание: роман в шести частях с эпилогом
    </label>
    <elementList rdf:parseType="Collection">
      <mainTitleElement>
        <elementValue>Преступление и наказание</elementValue>
      </mainTitleElement>
      <SubTitleElement>
        <elementValue>роман в шести частях с эпилогом</elementValue>
      </SubTitleElement>
    </elementList>
  </Title>

```

</titlePrincipal>

<identifier:isbn>978-5-486-02129-9 (В пер.)</identifier:isbn>

<subjectComplex xmlns="http://www.loc.gov/mods/rdf/v1#">

<ComplexSubject xmlns="http://www.loc.gov/mads/rdf/v1#">

<rdfs:label>

Филологические науки. Художественная литература -- Россия -- Русская литература -- 2-ая пол. 19 в. (до сер. 90-х гг.) -- Произведения художественной литературы -- Художественная проза -- Романы. Повести. Рассказы.

</rdfs:label>

<componentList rdf:parseType="Collection">

<Topic>

<label xmlns="http://www.w3.org/2000/01/rdf-schema#">

Филологические науки. Художественная литература -- Россия -- Русская литература -- 2-ая пол. 19 в. (до сер. 90-х гг.) -- Произведения художественной литературы -- Художественная проза -- Романы. Повести. Рассказы

</label>

<elementList rdf:parseType="Collection">

<TopicElement>

<elementValue>

Филологические науки. Художественная литература -- Россия -- Русская литература -- 2-ая пол. 19 в. (до сер. 90-х гг.) -- Произведения художественной литературы -- Художественная проза -- Романы. Повести. Рассказы

</elementValue>

</TopicElement>

</elementList>

</Topic>

</componentList>

</ComplexSubject>

</subjectComplex>

<modsrdf:classificationGroup>

```

<ClassificationGroup>
  <classificationGroupScheme>udc</classificationGroupScheme>
  <classificationGroupValue>821.161.1-31</classificationGroupValue>
</ClassificationGroup>
</modsrdf:classificationGroup>
<modsrdf:classificationGroup>
  <ClassificationGroup>
    <classificationGroupScheme>rubbk</classificationGroupScheme>
    <classificationGroupValue>III5(2=P)52-644</classificationGroupValue>
  </ClassificationGroup>
</modsrdf:classificationGroup>
<LanguageOfResource                                xmlns="http://www.loc.gov/mods/rdf/v1#"
rdf:resource="http://id.loc.gov/vocabulary/language#rus"/>
<locationOfResource xmlns="http://www.loc.gov/mods/rdf/v1#">
  <Location>
    <locationPhysicalLocation>RGB FB</locationPhysicalLocation>
    <locationShelfLocator>10 08-20/305</locationShelfLocator>
  </Location>
</locationOfResource>
<locationOfResource xmlns="http://www.loc.gov/mods/rdf/v1#">
  <Location>
    <locationPhysicalLocation>RGB ORF</locationPhysicalLocation>
  </Location>
</locationOfResource>
<placeOfOrigin xmlns="http://www.loc.gov/mods/rdf/v1#">
  <Geographic xmlns="http://www.loc.gov/mads/rdf/v1#">
    <label xmlns="http://www.w3.org/2000/01/rdf-schema#">ru</label>
    <elementList rdf:parseType="Collection">
      <GeographicElement>
        <elementValue>ru</elementValue>

```

```

</GeographicElement>
</elementList>
</Geographic>
</placeOfOrigin>
<placeOfOrigin xmlns="http://www.loc.gov/mods/rdf/v1#">
  <Geographic xmlns="http://www.loc.gov/mads/rdf/v1#">
    <label xmlns="http://www.w3.org/2000/01/rdf-schema#">Москва</label>
    <elementList rdf:parseType="Collection">
      <GeographicElement>
        <elementValue>Москва</elementValue>
      </GeographicElement>
    </elementList>
  </Geographic>
</placeOfOrigin>
<publisher xmlns="http://www.loc.gov/mods/rdf/v1#">
  <CorporateName xmlns="http://www.loc.gov/mads/rdf/v1#">
    <label xmlns="http://www.w3.org/2000/01/rdf-schema#">Мир книги</label>
    <elementList rdf:parseType="Collection">
      <FullNameElement>
        <elementValue>Мир книги</elementValue>
      </FullNameElement>
    </elementList>
  </CorporateName>
</publisher>
<publisher xmlns="http://www.loc.gov/mods/rdf/v1#">
  <CorporateName xmlns="http://www.loc.gov/mads/rdf/v1#">
    <label xmlns="http://www.w3.org/2000/01/rdf-schema#">Литература</label>
    <elementList rdf:parseType="Collection">
      <FullNameElement>
        <elementValue>Литература</elementValue>

```

```

    </FullNameElement>
  </elementList>
</CorporateName>
</publisher>
  <resourceDateIssued
                                xmlns="http://www.loc.gov/mods/rdf/v1#"
rdf:datatype="xsd:string">2008</resourceDateIssued>
  <issuance xmlns="http://www.loc.gov/mods/rdf/v1#">monographic</issuance>
  <modsrdf:statementOfResponsibility>Федор                      Михайлович
Достоевский</modsrdf:statementOfResponsibility>
  <physicalForm xmlns="http://www.loc.gov/mods/rdf/v1#">print</physicalForm>
  <physicalExtent  xmlns="http://www.loc.gov/mods/rdf/v1#">510,  [1]  с.  21
см</physicalExtent>
  <administrativeMedatata xmlns="http://www.loc.gov/mods/rdf/v1#">
    <AdministrativeMedatata xmlns="http://id.loc.gov/ontologies/RecordInfo#">
      <descriptionStandard>rcr</descriptionStandard>
      <recordContentSource>RuMoRKP</recordContentSource>
      <recordInfoRecordCreationDate
rdf:datatype="xsd:date">080422</recordInfoRecordCreationDate>
      <recordInfoRecordChangeDate
rdf:datatype="xsd:date">20080702124604.0</recordInfoRecordChangeDate>
      <recordIdentifier>003538908</recordIdentifier>
      <languageOfCataloging
rdf:resource="http://id.loc.gov/vocabulary/language#rus"/>
    </AdministrativeMedatata>
  </administrativeMedatata>
  <relatedSeries>
    <modsrdf:ModsResource
rdf:about="http://www.loc.gov/mods/rdf/v1#MODS123456">
      <titlePrincipal xmlns="http://www.loc.gov/mods/rdf/v1#">
        <Title xmlns="http://www.loc.gov/mads/rdf/v1#">

```

```

<label      xmlns="http://www.w3.org/2000/01/rdf-schema#">Бриллиантовая
коллекция</label>

<elementList rdf:parseType="Collection">
  <mainTitleElement>
    <elementValue>Бриллиантовая коллекция</elementValue>
  </mainTitleElement>
</elementList>
</Title>
</titlePrincipal>
</modsrdf:ModsResource>
</relatedSeries>
</modsrdf:ModsResource>
</rdf:RDF>

```

Свойства классов MODS/RDF

Свойства MODS/RDF	Соответствующий элемент MODS XML	Область	Значение
abstract	abstract	ModsResource	xsd:string
accessCondition	accessCondition		xsd:string
adminMetadata	recordInfo		ri:AdminMetadata
cartographics	subject/cartographics		Cartographics
class: {scheme}; e.g. class:lcc	classification		xsd:string
classificationGroup	classification		ClassificationGroup
classificationGroupScheme	ClassificationGroup	ClassificationGroup	xsd:string
classificationGroupValue			xsd:string
date	originInfo/dateOther	ModsResource	(annotation property)
dateCaptured	originInfo/dateCapture		

dateCapturedEnd	d	ModsResource	(annotation property)
dateCapturedStart			
dateCreated	originInfo/dateCreated		
dateCreatedEnd			
dateCreatedStart			
dateEnd	originInfo/dateOther		
dateIssued	originInfo/dateIssued		
dateIssuedEnd			
dateIssuedStart			
dateModified	originInfo/dateModifi ed		
dateModifiedEnd			
dateModifiedStart			
dateOfCopyright	originInfo/copyrightD ate		
dateOfCopyrightEnd			
dateOfCopyrightStart			
dateStart	originInfo/dateOther		
dateValid	originInfo/dateValid		
dateValidEnd			
dateValidStart			
digitalOrigin	physicalDescription/di gitalOrigin		xsd:string
edition	originInfo/edition		
extent	physicalDescription/e xtent		
form	physicalDescription/fo rm		
frequency	originInfo/frequency		
genre	genre		madsrdf:GenreForm

identifier: {type}; e.g. identifier:isbn			xsd:string
identifierGroup	identifier		IdentifierGroup
identifierGroupType		IdentifierGroup	xsd:string
identifierGroupValue		up	
issuance	originInfo/issuance	ModsResource	
languageOfResource	language	e	(annotation property)
LocationCopy	location/holdingSimple/Copy	Location	LocationCopy
locationCopyElectronicLocator	location/holdingSimple/copy/electronicLocator	LocationCopy	xsd:string
locationCopyForm	location/holdingSimple/copy/form		
locationCopyNote	location/holdingSimple/copy/note		
locationCopyShelfLocator	location/holdingSimple/copy/shelfLocator		
locationCopySublocation	location/holdingSimple/copy/sublocation		
locationOfResource	location	Location	
locationPhysical	location/physical		
locationShelfLocator	location/shelfLocator		
locationSublocation	location/sublocation		
locationUrl	location/url		

mediaType	physicalDescription/internetMediaType	ModsResource	
name	name		madsrdf:Name
namePrincipal			madsrdf:Name
noteGroup	note		NoteGroup
noteGroupType		NoteGroup	xsd:string
noteGroupValue			
noteUntyped		ModsResource	xsd:string
part	part	e	Part
partCaption	part/detail/caption	Part	xsd:string
partDate	part/date		(annotation property)
partDetailType	part/detail@type		xsd:string
partEnd	part/extent/end		xsd:string
partLevel	part/detail/level		xsd:positiveInteger
partList	part/extent/list		xsd:string
partTotal	part/extent/total		xsd:positiveInteger
partNumber	part/detail/number		xsd:string
partOrder	part@order		xsd:integer
partStart	part/extent/start		xsd:string
partText	part/text		
partTitle	part/detail/title		
partTotal	part/extent/list		
partType	part@type		
partUnit	part/extent@unit		
physicalDescriptionNote	physicalDescription/note	ModsResource	
reformattingQuality	physicalDescription/reformattingQuality		

relatedConstituent	relatedItem	ModResource
relatedFormat		
relatedHost		
relatedItem		
relatedOriginal		
relatedPreceding		
relatedReference		
relatedReferencedBy		
relatedReview		
relatedSeries		
relatedSucceeding		
relatedVersion		
statementOfResponsibility	note	xsd:string
subjectComplex	subject	madsrdf:ComplexSubject
subjectGenre		madsrdf:GenreForm
subjectGeographic		madsrdf:Geographic
subjectGeographicCode		madsrdf:GeographicCode
subjectHierarchicalGeographic		madsrdf:ComplexSubject
subjectName		madsrdf:Name
subjectOccupation		madsrdf:Occupation
subjectTemporal		madsrdf:Temporal
subjectTitle		madsrdf:Title
subjectTopic		madsrdf:Topic
tableOfContents	tableOfContents	xsd:string
targetAudience	targetAudience	

titlePrincipal	titleInfo	madsrdf:Title
titleUniform		

Приложение Г

Список библиографических записей, получаемых по протоколу OAI-PMH

```

<OAI-PMH                                xmlns="http://www.openarchives.org/OAI/2.0/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
<responseDate>2015-03-23T08:50:43Z</responseDate>
<request      verb="ListRecords"      from="2015-03-12"      until="2015-03-12"
metadataPrefix="marc21">http://aleph.nlr.ru/OAI</request>
<ListRecords>
  <record>
    <header>
      <identifier>oai:aleph.nlr.ru:NLR01-000889961</identifier>
      <timestamp>2015-03-12T02:05:22Z</timestamp>
      <setSpec>NLR01</setSpec>
    </header>
    <metadata>
      <marc:record                                xmlns:marc="http://www.loc.gov/MARC21/slim"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.loc.gov/MARC21/slim
http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd">
        <marc:leader>nam0 2200241 i 450</marc:leader>
        <marc:controlfield tag="001">RU\NLR\bibl\1239333</marc:controlfield>
        <marc:controlfield tag="005">20080923175140.0</marc:controlfield>
        <marc:datafield tag="010" ind1=" " ind2=" ">
          <marc:subfield code="a">978-5-486-02129-9</marc:subfield>
          <marc:subfield code="9">10000</marc:subfield>

```

```

</marc:datafield>
<marc:datafield tag="021" ind1=" " ind2=" ">
  <marc:subfield code="a">RU</marc:subfield>
  <marc:subfield code="b">2008-28354</marc:subfield>
  <marc:subfield code="9">1570</marc:subfield>
</marc:datafield>
<marc:datafield tag="021" ind1=" " ind2=" ">
  <marc:subfield code="a">RU</marc:subfield>
  <marc:subfield code="b">2008-28354</marc:subfield>
</marc:datafield>
<marc:datafield tag="100" ind1=" " ind2=" ">
  <marc:subfield code="a">20080609d2008 u y0rusy50 ca</marc:subfield>
</marc:datafield>
<marc:datafield tag="101" ind1="0" ind2=" ">
  <marc:subfield code="a">rus</marc:subfield>
</marc:datafield>
<marc:datafield tag="102" ind1=" " ind2=" ">
  <marc:subfield code="a">RU</marc:subfield>
</marc:datafield>
<marc:datafield tag="105" ind1=" " ind2=" ">
  <marc:subfield code="a">y |||||</marc:subfield>
</marc:datafield>
<marc:datafield tag="200" ind1="1" ind2=" ">
  <marc:subfield code="a">Преступление и наказание</marc:subfield>
  <marc:subfield code="e">роман в шести частях с эпилогом</marc:subfield>
  <marc:subfield code="f">Ф.М. Достоевский</marc:subfield>
  <marc:subfield code="g">[вступ. ст. А. Кирпичникова</marc:subfield>
  <marc:subfield code="g">коммент. Г. Когана и Г.
Фридлендера]</marc:subfield>
</marc:datafield>

```

```

<marc:datafield tag="210" ind1=" " ind2=" ">
  <marc:subfield code="a">Москва</marc:subfield>
  <marc:subfield code="c">Мир книги</marc:subfield>
  <marc:subfield code="c">Литература</marc:subfield>
  <marc:subfield code="d">2008</marc:subfield>
</marc:datafield>

<marc:datafield tag="215" ind1=" " ind2=" ">
  <marc:subfield code="a">510,[1] с.</marc:subfield>
  <marc:subfield code="d">21</marc:subfield>
</marc:datafield>

<marc:datafield tag="225" ind1="1" ind2=" ">
  <marc:subfield code="a">Бриллиантовая коллекция</marc:subfield>
</marc:datafield>

<marc:datafield tag="686" ind1=" " ind2=" ">
  <marc:subfield code="a">III6(2=P)52-4</marc:subfield>
  <marc:subfield code="v">LBC/SL</marc:subfield>
  <marc:subfield code="2">rubbk</marc:subfield>
</marc:datafield>

<marc:datafield tag="700" ind1=" " ind2="1">
  <marc:subfield code="3">RU\NLR\auth\77838</marc:subfield>
  <marc:subfield code="a">Достоевский</marc:subfield>
  <marc:subfield code="b">Ф. М.</marc:subfield>
  <marc:subfield code="f">1821-1881</marc:subfield>
  <marc:subfield code="g">Федор Михайлович</marc:subfield>
  <marc:subfield code="4">070</marc:subfield>
</marc:datafield>

<marc:datafield tag="801" ind1=" " ind2="0">
  <marc:subfield code="a">RU</marc:subfield>
  <marc:subfield code="b">NLR</marc:subfield>
  <marc:subfield code="c">20080609</marc:subfield>

```

```

    <marc:subfield code="g">RCR</marc:subfield>
  </marc:datafield>
  <marc:datafield tag="801" ind1=" " ind2="1">
    <marc:subfield code="a">RU</marc:subfield>
    <marc:subfield code="b">NLR</marc:subfield>
    <marc:subfield code="c">20080609</marc:subfield>
  </marc:datafield>
  <marc:datafield tag="801" ind1=" " ind2="2">
    <marc:subfield code="a">RU</marc:subfield>
    <marc:subfield code="b">NLR</marc:subfield>
    <marc:subfield code="c">20150729</marc:subfield>
    <marc:subfield code="2">rusmarc</marc:subfield>
  </marc:datafield>
  <marc:datafield tag="899" ind1=" " ind2=" ">
    <marc:subfield code="a">NLR</marc:subfield>
    <marc:subfield code="j">2008-3/17379</marc:subfield>
  </marc:datafield>
</marc:record>
</metadata>
</record>
<record>
  <header>
    <identifier>oai:aleph.nlr.ru:NLR01-001967333</identifier>
    <timestamp>2015-03-12T02:05:23Z</timestamp>
    <setSpec>NLR01</setSpec>
  </header>
  <metadata>
    <marc:record
      xmlns:marc="http://www.loc.gov/MARC21/slim"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"

```

xsi:schemaLocation="http://www.loc.gov/MARC21/slim

http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd">

<marc:leader>nam2 2200205 i 450</marc:leader>

<marc:controlfield tag="001">RU\NLR\bibl\1978596</marc:controlfield>

<marc:controlfield tag="005">20130402110417.0</marc:controlfield>

<marc:datafield tag="010" ind1=" " ind2=" ">

<marc:subfield code="9">100</marc:subfield>

</marc:datafield>

<marc:datafield tag="100" ind1=" " ind2=" ">

<marc:subfield code="a">20121212d2010 u y0rusy50 ca</marc:subfield>

</marc:datafield>

<marc:datafield tag="101" ind1="0" ind2=" ">

<marc:subfield code="a">rus</marc:subfield>

</marc:datafield>

<marc:datafield tag="102" ind1=" " ind2=" ">

<marc:subfield code="a">RU</marc:subfield>

</marc:datafield>

<marc:datafield tag="105" ind1=" " ind2=" ">

<marc:subfield code="a">a |||||</marc:subfield>

</marc:datafield>

<marc:datafield tag="200" ind1="0" ind2=" ">

<marc:subfield code="a">... год 2009</marc:subfield>

<marc:subfield code="f">

Деп. культуры и нац. политики Кемер. обл., Гос. учреждение культуры "Кемер. обл. науч. б-ка им. В.Д. Федорова", Отд. прогнозирования и развития библиотеч. дела

</marc:subfield>

<marc:subfield code="g">[отв. ред.: Майнгардт Н.Л.]</marc:subfield>

</marc:datafield>

<marc:datafield tag="210" ind1=" " ind2=" ">

<marc:subfield code="a">Кемерово</marc:subfield>

```

<marc:subfield code="c">[б. и.]</marc:subfield>
<marc:subfield code="d">2010</marc:subfield>
</marc:datafield>
<marc:datafield tag="215" ind1=" " ind2=" ">
  <marc:subfield code="a">134 с.</marc:subfield>
  <marc:subfield code="c">ил.</marc:subfield>
</marc:datafield>
<marc:datafield tag="461" ind1=" " ind2="0">
  <marc:subfield code="1">001RU\NLR\bibl\912723</marc:subfield>
  <marc:subfield code="1">2001</marc:subfield>
  <marc:subfield code="a">

```

Ежегодный доклад о деятельности государственных и муниципальных библиотек Кемеровской области: ...

```

</marc:subfield>
<marc:subfield code="v">... год 2009</marc:subfield>
</marc:datafield>
<marc:datafield tag="712" ind1="0" ind2="2">
  <marc:subfield code="3">RU\NLR\auth\8810230525</marc:subfield>
  <marc:subfield code="a">

```

Кемеровская областная научная библиотека им. В. Д. Федорова

```

</marc:subfield>
<marc:subfield code="b">

```

Отдел прогнозирования и развития библиотечного дела

```

</marc:subfield>
<marc:subfield code="4">570</marc:subfield>
</marc:datafield>
<marc:datafield tag="801" ind1=" " ind2="0">
  <marc:subfield code="a">RU</marc:subfield>
  <marc:subfield code="b">NLR</marc:subfield>
  <marc:subfield code="c">20121212</marc:subfield>

```

```

    <marc:subfield code="g">RCR</marc:subfield>
  </marc:datafield>
  <marc:datafield tag="801" ind1=" " ind2="1">
    <marc:subfield code="a">RU</marc:subfield>
    <marc:subfield code="b">NLR</marc:subfield>
    <marc:subfield code="c">20121212</marc:subfield>
  </marc:datafield>
  <marc:datafield tag="801" ind1=" " ind2="2">
    <marc:subfield code="a">RU</marc:subfield>
    <marc:subfield code="b">NLR</marc:subfield>
    <marc:subfield code="c">20150312</marc:subfield>
    <marc:subfield code="2">rusmarc</marc:subfield>
  </marc:datafield>
  <marc:datafield tag="899" ind1=" " ind2=" ">
    <marc:subfield code="a">NLR</marc:subfield>
    <marc:subfield code="j">Б 6020/Е-365</marc:subfield>
  </marc:datafield>
</marc:record>
</metadata>
</record>
<resumptionToken>201503120211275201503129999999NLR01:OAI-FULL-
MARC</resumptionToken>
</ListRecords>
</OAI-PMH>

```

Приложение Д

XSLT-шаблон преобразования библиографических записей

```

<xsl:stylesheet version="1.0"
  xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
  xmlns:mods="http://www.loc.gov/mods/v3"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:foaf="http://xmlns.com/foaf/0.1/"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:dcterms="http://purl.org/dc/terms/"
  xmlns:owl="http://www.w3.org/2002/07/owl#">
  <xsl:output
    method="xml"
    indent="yes"
    encoding="UTF-8"
    omit-xml-declaration="yes" />
  <xsl:template match="/mods:mods">
    <rdf:RDF>
      <xsl:apply-templates select="mods:titleInfo"/>
      <xsl:apply-templates select="mods:name"/>
    </rdf:RDF>
  </xsl:template>
  <xsl:template match="mods:titleInfo">
    <rdf:Description>
      <xsl:choose>
        <xsl:when test="../mods:recordInfo/mods:recordIdentifier">
          <xsl:call-template name="aboutLink">
            <xsl:with-param name="path" select="'title'" />

```

```

        <xsl:with-param                                name="identifier"
select="../mods:recordInfo/mods:recordIdentifier" />
    </xsl:call-template>
</xsl:when>
<xsl:otherwise>
    <xsl:attribute name="rdf:about">
        <xsl:value-of select="concat('http://title#', mods:title)" />
    </xsl:attribute>
</xsl:otherwise>
</xsl:choose>
<xsl:variable name="title">
    <xsl:value-of select="mods:nonSort"/>
    <xsl:if test="mods:nonSort">
        <xsl:text> </xsl:text>
    </xsl:if>
    <xsl:value-of select="mods:title"/>
    <xsl:if test="mods:subTitle">
        <xsl:text> </xsl:text>
        <xsl:value-of select="mods:subTitle"/>
    </xsl:if>
    <xsl:if test="mods:partNumber">
        <xsl:text> </xsl:text>
        <xsl:value-of select="mods:partNumber"/>
    </xsl:if>
    <xsl:if test="mods:partName">
        <xsl:text> </xsl:text>
        <xsl:value-of select="mods:partName"/>
    </xsl:if>
</xsl:variable>
<dc:title>

```

```

    <xsl:value-of select="normalize-space($title)"/>
</dc:title>
<xsl:for-each select=" ../mods:name">
  <xsl:choose>
    <xsl:when test="(mods:role/mods:roleTerm = 'aut')
    or
    (mods:role/mods:roleTerm = 'автор')
    or
    (mods:role/mods:roleTerm = 'creator')
    or
    (mods:role/mods:roleTerm = 'cre')
    ">
    <dc:creator>
      <xsl:choose>
        <xsl:when test=" ../mods:recordInfo/mods:recordIdentifier">
          <xsl:call-template name="resourceLink">
            <xsl:with-param name="path" select=""author"" />
            <xsl:with-param name="identifier"
select=" ../mods:recordInfo/mods:recordIdentifier" />
          </xsl:call-template>
        </xsl:when>
        <xsl:otherwise>
          <xsl:attribute name="rdf:resource">
            <xsl:value-of select="concat('http://author#', mods:namePart)"/>
          </xsl:attribute>
        </xsl:otherwise>
      </xsl:choose>
    </dc:creator>
  </xsl:when>
  <xsl:otherwise>

```

```

<dc:contributor>
  <xsl:choose>
    <xsl:when test="../mods:recordInfo/mods:recordIdentifier">
      <xsl:call-template name="resourceLink">
        <xsl:with-param name="path" select="'author'" />
        <xsl:with-param name="identifier"
select="../mods:recordInfo/mods:recordIdentifier" />
      </xsl:call-template>
    </xsl:when>
    <xsl:otherwise>
      <xsl:attribute name="rdf:resource">
        <xsl:value-of select="concat('http://author#', mods:namePart)"/>
      </xsl:attribute>
    </xsl:otherwise>
  </xsl:choose>
</dc:contributor>
</xsl:otherwise>
</xsl:choose>
</xsl:for-each>
<xsl:apply-templates select="../mods:genre"/>
<xsl:apply-templates select="../mods:typeOfResource"/>
<xsl:apply-templates select="../mods:originInfo"/>
<xsl:apply-templates select="../mods:language"/>
<xsl:apply-templates select="../mods:physicalDescription"/>
<xsl:apply-templates select="../mods:abstract"/>
<xsl:apply-templates select="../mods:tableOfContents"/>
<xsl:apply-templates select="../mods:note"/>
<xsl:apply-templates select="../mods:subject"/>
<xsl:apply-templates select="../mods:identifier"/>
<xsl:apply-templates select="../mods:location"/>

```

```

    <xsl:apply-templates select="../mods:accessCondition"/>
  </rdf:Description>
</xsl:template>
<xsl:template match="mods:name">
  <rdf:Description>
    <xsl:choose>
      <xsl:when test="../mods:recordInfo/mods:recordIdentifier">
        <xsl:call-template name="aboutLink">
          <xsl:with-param name="path" select="'author'" />
          <xsl:with-param name="identifier"
select="../mods:recordInfo/mods:recordIdentifier" />
        </xsl:call-template>
      </xsl:when>
      <xsl:otherwise>
        <xsl:attribute name="rdf:about">
          <xsl:value-of select="concat('http://author#', mods:namePart)" />
        </xsl:attribute>
      </xsl:otherwise>
    </xsl:choose>
    <xsl:variable name="name">
      <xsl:for-each select="mods:namePart[not(@type)]">
        <xsl:value-of select="."/>
        <xsl:text> </xsl:text>
      </xsl:for-each>
    </xsl:variable>
    <foaf:name>
      <xsl:value-of select="normalize-space($name)"/>
    </foaf:name>
    <xsl:if test="mods:namePart[@type='given']">
      <foaf:givenName>

```

```

    <xsl:value-of select="mods:namePart[@type='given']"/>
  </foaf:givenName>
</xsl:if>
<xsl:if test="mods:namePart[@type='family']">
  <foaf:familyName>
    <xsl:value-of select="mods:namePart[@type='family']"/>
  </foaf:familyName>
</xsl:if>
<xsl:if test="mods:displayForm">
  <foaf:name>
    <xsl:value-of select="mods:displayForm"/>
  </foaf:name>
</xsl:if>
<xsl:if test="@type='personal'">
  <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Person"/>
</xsl:if>
<xsl:if test="@type='corporate'">
  <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Organization"/>
</xsl:if>
</rdf:Description>
</xsl:template>
<xsl:template match="mods:subject">
  <xsl:if test="mods:topic | mods:occupation">
    <dc:subject>
      <xsl:for-each select="mods:topic | mods:occupation">
        <xsl:value-of select="."/>
        <xsl:if test="position() != last()">--</xsl:if>
      </xsl:for-each>
    </dc:subject>
  </xsl:if>

```

```

<xsl:for-each select="mods:titleInfo">
  <dc:subject>
    <xsl:for-each select="child::*">
      <xsl:value-of select="."/>
      <xsl:if test="following-sibling::*"><xsl:text> </xsl:text></xsl:if>
    </xsl:for-each>
  </dc:subject>
</xsl:for-each>
<xsl:for-each select="mods:geographic">
  <dc:coverage>
    <xsl:value-of select="."/>
  </dc:coverage>
</xsl:for-each>
<xsl:for-each select="mods:hierarchicalGeographic">
  <dc:coverage>
    <xsl:for-each
select="mods:continent|mods:country|mods:province|mods:region|mods:state|mods:terri
tory|mods:county|mods:city|mods:island|mods:area">
      <xsl:value-of select="."/>
      <xsl:if test="position()=last()">--</xsl:if>
    </xsl:for-each>
  </dc:coverage>
</xsl:for-each>
<xsl:for-each select="mods:cartographics/*">
  <dc:coverage>
    <xsl:value-of select="."/>
  </dc:coverage>
</xsl:for-each>
<xsl:if test="mods:temporal">
  <dc:coverage>

```

```

<xsl:for-each select="mods:temporal">
  <xsl:value-of select="."/>
  <xsl:if test="position()=last()">-</xsl:if>
</xsl:for-each>
</dc:coverage>
</xsl:if>
<xsl:if test="*[1][local-name()='topic'] and *[local-name()!='topic']">
  <dc:subject>
    <xsl:for-each      select="*[local-name()!='cartographics'      and      local-
name()!='geographicCode' and local-name()!='hierarchicalGeographic'] ">
      <xsl:value-of select="."/>
      <xsl:if test="position()=last()">--</xsl:if>
    </xsl:for-each>
  </dc:subject>
</xsl:if>
</xsl:template>
<xsl:template match="mods:abstract | mods:tableOfContents | mods:note">
  <dc:description>
    <xsl:value-of select="."/>
  </dc:description>
</xsl:template>
<xsl:template match="mods:genre">
  <xsl:choose>
    <xsl:when test="@authority='dct'">
      <dc:type>
        <xsl:value-of select="."/>
      </dc:type>
    </xsl:when>
    <xsl:otherwise>
      <dc:type>

```

```

    <xsl:value-of select="."/>
  </dc:type>
  <xsl:apply-templates select="mods:typeOfResource"/>
</xsl:otherwise>
</xsl:choose>
</xsl:template>
<xsl:template match="mods:typeOfResource">
  <xsl:if test="@collection='yes'">
    <dc:type>Collection</dc:type>
  </xsl:if>
  <xsl:if test=".='software' and ../mods:genre='database'">
    <dc:type>Dataset</dc:type>
  </xsl:if>
  <xsl:if test=".='software' and ../mods:genre='online system or service'">
    <dc:type>Service</dc:type>
  </xsl:if>
  <xsl:if test=".='software'">
    <dc:type>Software</dc:type>
  </xsl:if>
  <xsl:if test=".='cartographic material'">
    <dc:type>Image</dc:type>
  </xsl:if>
  <xsl:if test=".='multimedia'">
    <dc:type>InteractiveResource</dc:type>
  </xsl:if>
  <xsl:if test=".='moving image'">
    <dc:type>MovingImage</dc:type>
  </xsl:if>
  <xsl:if test=".='three dimensional object'">
    <dc:type>PhysicalObject</dc:type>

```

```

</xsl:if>
<xsl:if test="starts-with(.,'sound recording')">
  <dc:type>Sound</dc:type>
</xsl:if>
<xsl:if test=".='still image'">
  <dc:type>StillImage</dc:type>
</xsl:if>
<xsl:if test=".='text'">
  <dc:type>Text</dc:type>
</xsl:if>
<xsl:if test=".='notated music'">
  <dc:type>Text</dc:type>
</xsl:if>
</xsl:template>
<xsl:template match="mods:physicalDescription">
  <xsl:for-each select="mods:extent | mods:form | mods:internetMediaType">
    <dc:format>
      <xsl:variable name="unit" select="@unit"/>
      <xsl:if test="@unit"><xsl:value-of select="$unit"/>: </xsl:if><xsl:value-of
select="."/>
    </dc:format>
  </xsl:for-each>
</xsl:template>
<xsl:template match="mods:identifier">
  <dc:identifier>
    <xsl:choose>
      <xsl:when test="contains(.,':')">
        <xsl:value-of select="."/>
      </xsl:when>
      <xsl:when test="@type">

```

```

    <xsl:value-of select="@type"/>: <xsl:value-of select="."/>
  </xsl:when>

  <xsl:when test="contains('isbn issn uri doi lccn uri', lower-case(.))">
    <xsl:value-of select="."/>: <xsl:value-of select="."/>
  </xsl:when>

  <xsl:otherwise>
    <xsl:value-of select="."/>
  </xsl:otherwise>
</xsl:choose>
</dc:identifier>
</xsl:template>
<xsl:template match="mods:location">
  <xsl:for-each select="mods:url">
    <dc:identifier>
      <xsl:value-of select="."/>
    </dc:identifier>
  </xsl:for-each>
</xsl:template>
<xsl:template match="mods:language">
  <dc:language>
    <xsl:value-of select="child::*"/>
  </dc:language>
</xsl:template>
<xsl:template match="mods:accessCondition">
  <dc:rights>
    <xsl:value-of select="."/>
  </dc:rights>
</xsl:template>
<xsl:template match="mods:originInfo">
  <xsl:apply-templates select="*[@point='start']"/>

```

```

<xsl:for-each select="mods:publisher">
  <dc:publisher>
    <xsl:value-of select="."/>
  </dc:publisher>
</xsl:for-each>
</xsl:template>
<xsl:template match="mods:dateIssued | mods:dateCreated | mods:dateCaptured">
  <dc:date>
    <xsl:choose>
      <xsl:when test="@point='start'">
        <xsl:variable name="dateName" select="local-name(.)"/>
        <xsl:value-of select="."/>
        <xsl:text>-</xsl:text>
        <xsl:value-of select="../*[local-name()=$dateName][@point='end']"/>
      </xsl:when>
      <xsl:otherwise>
        <xsl:value-of select="."/>
      </xsl:otherwise>
    </xsl:choose>
  </dc:date>
</xsl:template>
<xsl:template name="aboutLink" match="mods:recordInfo">
  <xsl:param name="path"/>
  <xsl:param name="identifier"/>
  <xsl:attribute name="rdf:about">
    <xsl:value-of select="concat('http://www.rusneb.ru/', $path, '/', replace($identifier,
'\', '/'))" />
  </xsl:attribute>
</xsl:template>
<xsl:template name="resourceLink" match="mods:recordInfo">

```

```

<xsl:param name="path"/>
<xsl:param name="identifier"/>
<xsl:attribute name="rdf:resource">
  <xsl:value-of select="concat('http://www.rusneb.ru/', $path, '/', replace($identifier,
'\\', '/'))" />
</xsl:attribute>
</xsl:template>
<xsl:template match="mods:temporal[@point='start'] ">
  <xsl:value-of select="."/>-<xsl:value-of select="../mods:temporal[@point='end']"/>
</xsl:template>
<xsl:template match="mods:temporal[@point!='start' and @point!='end']">
  <xsl:value-of select="."/>
</xsl:template>
</xsl:stylesheet>

```

Пример связанных данных в формате RDF/XML

```

<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:mods="http://www.loc.gov/mods/v3"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:dcterms="http://purl.org/dc/terms/"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:foaf="http://xmlns.com/foaf/0.1/"
  xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description rdf:about="http://www.rusneb.ru/title/RU/NLR/bibl/1239333">
    <dc:title>Преступление и наказание роман в шести частях с эпилогом</dc:title>
    <dc:contributor
rdf:resource="http://www.rusneb.ru/author/RU/NLR/bibl/1239333"/>
    <dc:type>Text</dc:type>
    <dc:publisher>Мир книги</dc:publisher>

```

```

<dc:publisher>Литература</dc:publisher>
<dc:language>rus</dc:language>
<dc:format>print</dc:format>
<dc:format>510, [1] с. 21 см</dc:format>
<dc:description>Федор Михайлович Достоевский</dc:description>
<dc:subject>Филологические науки. Художественная литература -- Россия --
Русская литература -- 2-ая пол. 19 в. (до сер. 90-х гг.) -- Произведения
художественной литературы -- Художественная проза -- Романы. Повести.
Рассказы</dc:subject>
<dc:identifier>isbn: 978-5-486-02129-9 (В пер.)</dc:identifier>
<rdfs:seeAlso rdf:resource="http://lcn.loc.gov/91053120"
  rdfs:label="Crime and punishment : a novel in six parts with epilogue /"/>
<rdfs:seeAlso rdf:resource="http://lcn.loc.gov/92056354"
  rdfs:label="Crime and punishment : a novel in six parts with epilogue /"/>
</rdf:Description>
<rdf:Description rdf:about="http://www.rusneb.ru/author/RU/NLR/bibl/1239333">
  <foaf:name>Достоевский, Федор Михайлович</foaf:name>
  <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Person"/>
</rdf:Description>
</rdf:RDF>

```