На правах рукописи

Шорин Олег Николаевич

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

05.13.11 – математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Санкт-Петербург – 2017

Работа выполнена в Федеральном государственном бюджетном учреждении науки Санкт-Петербургском институте информатики и автоматизации Российской академии наук (СПИИРАН).

Научный руководитель: доктор физико-математических наук, профессор
Серебряков Владимир Алексеевич

Официальные оппоненты: **Горбунов-Посадов Михаил Михайлович**
доктор физико-математических наук, профессор, начальник отдела Института прикладной математики им. М.В. Келдыша

Калёнов Николай Евгеньевич
доктор технических наук, профессор, директор Федерального государственного бюджетного учреждения науки Библиотека по естественным наукам РАН

Ведущая организация: Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный университет имени М.В. Ломоносова».

Защита диссертации состоится «29» марта 2017 г. в 16 ч. 00 мин. на заседании диссертационного совета Д 002.073.02 при Федеральном исследовательском центре «Информатика и управление» Российской академии наук по адресу: 119333, Москва, Вавилова, д.44, кор.2.

С диссертацией можно ознакомиться в библиотеке Федерального исследовательского центра «Информатика и управление» Российской академии наук по адресу г.Москва, ул. Вавилова, д.44, кор. 2 и на сайте www.frccsc.ru.

Автореферат разослан “___”_____2017 г.

Ученый секретарь
диссертационного совета Р.В. Разумчик

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. Представлением библиографических записей в виде связанных данных занимаются ведущие каталогизаторы и специалисты информационных технологий. В интернете реализуется проект открытых связанных данных – Linked Open Data (LOD), 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.

На 78-ом Всемирном библиотечном конгрессе Международной федерации библиотечных ассоциаций IFLA, который состоялся в 2012 году в Хельсинки, были озвучены следующие преимущества от публикации библиотечных данных в связанном виде:

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

При публикации данных, собранных из различных источников, неизбежно возникают вопросы интеграции: выявления дублетных записей и их слияния. В мире существует всего несколько проектов по решению комплексной проблемы интеграции данных из различных библиотек с выявлением дублетных записей и последующим их слиянием, преобразованием в связанные данные и публикацией в LOD. Методы, применяемые в этих проектах, не могут быть использованы в случае большого объема разнородных данных, поскольку существующие подходы основаны на использовании единого формата представления данных и неавтоматизированном механизме связывания данных.

Целью диссертационной работы является расширение возможностей интеграции библиографических записей в открытое семантическое пространство.

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

Реализация поставленной цели предполагает решение следующих **подзадач**:

- Анализ существующих решений в области интеграции библиографических записей в открытое семантическое пространство.
- Разработка совместимой с уже существующими онтологии предметной области с учетом полноты представленной в библиографических записях информации.

- Разработка алгоритмов, обеспечивающих установление близости текстовых полей библиографических записей.
- Разработка системы интеграции библиографических записей, позволяющей формировать, хранить и предоставлять доступ к данным с использованием принципов Linked Open Data.

Результаты, выносимые на защиту:

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 отличии 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 исследования. В работе применялись методы сравнительного анализа, моделирования, классификации, непараметрической статистики, сопоставления строк, связывания записей, а также методы анализа, синтеза и тестирования 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 дублетные записи 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.

Реализация и внедрение результатов работы. Разработанная в диссертации система семантического связывания библиографических записей 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».

Личный вклад. Выносимые на защиту результаты получены соискателем лично. В опубликованных совместных работах постановка и исследование задач осуществлялись совместными усилиями соавторов при непосредственном участии соискателя.

Апробация работы. Основные положения диссертации изложены в 10 публикациях. 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.), EMC 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.).

Публикации. По материалам диссертации опубликовано 10 работ, из них 2 статьи в изданиях, входящих в перечень ВАК, 3 статьи в сборниках трудов конференций.

Структура и объем диссертационной работы. Диссертация состоит из введения, четырех глав и заключения. Каждая глава завершается выводами. Полный объем диссертации составляет 183 страницы. Список литературы содержит 156 наименований. В диссертации 7 рисунков, 12 таблиц, 1 график, приводится 5 приложений. Объем приложений составляет 60 страниц.

СОДЕРЖАНИЕ РАБОТЫ

Во введении аргументируется актуальность исследований, формулируются цель и задачи работы, перечисляются используемые методы исследования, обосновывается научная новизна и практическая значимость полученных результатов, приводятся сведения о результатах внедрения и использования.

Первая глава посвящена обзору предпосылок возникновения семантической паутины, анализу существующих решений по семантической интеграции библиографических записей. В главе осуществляется постановка задачи и описывается логическая схема исследования.

В разделе 1.1 приводится краткая история возникновения и развития Всемирной паутины – World Wide Web. Сделан анализ развития средств для ввода информации на сайты: визуальные редакторы, комментарии, блоги, социальные сети.

В 2006 году Тим Бернерс-Ли сформулировал основные принципы связанных данных - надстройки над существующим интернетом, которая позволяет автоматизированным системам извлекать информацию, анализировать её, устанавливать взаимосвязи и генерировать новую информацию.

Для реализации этих принципов было предложено использовать модель представления данных RDF (Resource Description Framework), которая пригодна для машинной обработки. Структурно выражения в RDF представляют собой триплеты, состоящие из субъекта, предиката и объекта. Выражение RDF-

триплета означает, что отношение, указанное предикатом, связывает предметы, обозначенные как субъект и объект.

В заключении раздела приводится обзор различных способов сериализации RDF: RDF/XML, RDFa, Turtle, N-Triples, RDF/JSON. Для каждого способа сериализации произведен анализ преимуществ и недостатков, а также описаны наиболее подходящие сферы применения.

В подразделе 1.1.1 показано, что для выражения семантики абстрактной модели RDF недостаточно. Необходимо использовать словари, таксономии и онтологии. В качестве примеров языков для составления этих сущностей приводятся RDFS, SKOS и OWL. Для каждого из этих трёх языков приведены примеры использования, а также основные отличия их друг от друга. В результате исследования делается вывод, что возможность использования терминов из уже существующих словарей является одним из основополагающих механизмов связанных данных. Поскольку в интернет сообществе уже разработан ряд словарей, которые описывают широко распространенные понятия, декларируется, что по возможности необходимо использовать уже существующие словари и онтологии.

В разделе 1.2 осуществляется обзор существующих проектов, предоставляющих доступ к библиографической информации с использованием принципов, предложенных Тимом Бернерс-Ли. Среди основных поставщиков библиографических данных в LOD выделяются 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.

В подразделе 1.2.1 подробно рассматривается проект 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.

Поставщики данных в Europeana обязаны сопровождать их описаниями, которые соответствуют единой модели – Europeana’s Semantic Elements (ESE). За основу ESE взят словарь Dublin Core Metadata Initiative (DCMI) Metadata Terms, который был дополнен 14 элементами, необходимыми для нужд Europeana. В разделе подробно рассматривается строение этой единой модели.

В 2011 году для описания всего многообразия культурных ценностей, хранящихся в Europeana, были начаты работы по созданию нового словаря – Europeana Data Model (EDM). В разделе также подробно рассматривается строение модели EDM.

В подразделе 1.2.2 рассматриваются исследования, проводимые в Библиотеке Конгресса США. Приводится обзор развития формата MARC (Machine-Readable Cataloging) с последующей эволюцией в схему представления записей в виде XML-файла – MARCXML, а также схем описания метаданных MODS и MADS.

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. Для онтологии MODS/RDF приведены основные классы. В приложении приведены примеры библиографической записи в формате MARC21, MARCXML, MODS и MODS/RDF, а также свойства классов MODS/RDF.

В подразделе 1.2.3 подробно рассматривается 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. Схема метаданных в ЕНИП состоит из 3 типов профилей:

- Базовый профиль, внутрь которого входит информация о предметных областях, применимых для всех учреждений РАН.
- Вспомогательный профиль, содержащий элементы, применимые не только к объектам научного мира.
- Специализированные профили, которые содержат в себе элементы, используемые в специализированных научных сообществах.

В разделе подробно рассматривается строение каждого профиля схемы метаданных.

В разделе 1.3 осуществляется постановка задачи и приводится логическая схема исследования. Показано, что 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, полученных из библиографических записей. Делается вывод, что в настоящий момент созрела необходимость решения следующих задач:

- Разработать совместимую с уже существующими онтологию предметной области с учетом полноты представленной в библиографических записях библиотек России информации. Онтология должна быть спроектирована таким образом, чтобы количество пустых данных было минимальным вне зависимости от того, по каким правилам каталогизации была создана библиографическая запись.
- 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.

- 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, формулируется структурная схема исследования.

Во второй главе проводится 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.

В разделе 2.1 анализируются протоколы для обмена библиографической информацией, реализованные в современных автоматизированных библиотечных информационных системах (АБИС). Рассматривается история развития протоколов Z39.50 и OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting). В разделе показано, что протокол Z39.50 используется для перекрестного поиска информации в многочисленных источниках, а протокол OAI-PMH - для автоматического сбора метаданных из различных АБИС и аккумуляции их на центральном сервере.

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.

Для осуществления сбора библиографических записей из библиотек созданы модули интеграции для автоматизированных библиотечных интегрированных систем, используемых в библиотеках: Aleph, Ирбис, MarcSQL и OPAC-Global.

В разделе приводится синтаксис запроса, посылаемого центральным порталом раз в сутки библиотекам России. В приложении приведен пример ответа в формате MARCXML, отсылаемого автоматизированной системой библиотеки на запрос из центрального узла системы.

В разделе 2.2 описывается алгоритм выявления дублетных библиографических записей с последующим их слиянием.

Вопросы интеграции библиографических записей из различных источников давно находятся в фокусе внимания ученых. Одними из основоположников этого направления являются Ivan P. Feleggi и Alan B. Sunter, которые разработали математическую модель выявления дублетных записей из двух различных множеств.

Описанная ими математическая модель имеет ряд недостатков, самым существенным из которых является необходимость вычисления функции сравнения для всех пар библиографических записей из двух множеств. В случае использования огромных массивов записей вычисление функции сравнения для всех возможных пар становится практически невыполнимым, поскольку сложность алгоритма в данном случае равна $O(n^2)$.

В работе Jeremy A. Hylton предлагается 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.

В диссертационной работе описывается алгоритм для выявления дублетных записей, который использует модифицированный подход Jeremy A. Hylton. Алгоритм состоит из 4 этапов. На первом этапе происходит нормализация библиографических записей:

- Записи конвертируются из формата MARCXML в MODS с использованием XSLT-шаблона.
- Последовательности пробельных символов заменяются на один пробел.
- Все буквы приводятся к нижнему регистру.
- Общепризнанные 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 множеству таких чисел в сбалансированном дереве осуществляется со сложностью $O(\log n)$.

После детерминированного поиска дублетов происходит вероятностный поиск. 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, достаточно сравнить только те записи, в которых «Год издания» либо совпадает, либо отсутствует. Данная оптимизация значительно снижает количество операций сравнения на втором этапе работы вероятностного алгоритма.

Для разбиения библиографических записей на кластеры используется функция хеширования SimHash, входящая в семейство locality-sensitive hashing

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

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

Основным свойством функций из данного семейства является то, что при незначительном изменении аргумента результат функции хеширования изменяется незначительно.

Шаги алгоритма вычисления хеша следующие:

1. Исходная строка разбивается на биграммы – 2-х буквенные сочетания, в результате чего 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. В разработанной системе для генерация 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 из 32 бит. Используется следующее соответствие элементов массива и хеша: если элемент массива больше нуля, то соответствующий бит равняется единице, иначе нулю.

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, так как если хеши, полученные из строк полей «Автор» и «Название», рассматривать как один, то для каждой пары хешей необходимо произвести 2^{64} операций побитового сравнения.

Для снижения количества сравнений используется метод разбиения хеша на равные части. Поиск близких хешей осуществляется с некоторой оптимизацией: среди всех 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 попадает 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.

В диссертационной работе 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 $1 \le i < j \le 4$ и $5 \le p < q \le 8$.

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 с использованием биграмм. Разбив поля «Автор» и «Название» на множество биграмм, вычисляется мера Жаккара, которая представляет собой соотношение количества совпадающих элементов этих множеств к суммарному количеству элементов в этих множествах:

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

В диссертационной работе доказываются две теоремы.

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

Наиболее затратной операцией при подсчете меры Жаккара является подсчет мощности пересечения двух множеств. Для оптимизации этого подсчета 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.

Имея отсортированные списки биграмм, используется Теорема №2 для того, чтобы не проходить полностью списки для некоторых пар.

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 $r = \lfloor (1-k) * (l(S_1) - 1) \rfloor + 1$ такое что, если $\forall p, q \le r$ и строки S_2 выполнено условие $Bi(S_1)[p] \ne Bi(S_2)[q]$, то для меры Жаккара двух множеств, состоящих из биграмм этих строк, выполнено неравенство $J(Bi(S_1),Bi(S_2))<k$.

Используя эти теоремы, реализуется дополнительная оптимизация. Для записей, претендующих на совпадение, пары строк из полей «Автор» и

«Название» проверяются на удовлетворение условию теорем при k, равном пороговому значению. Если пара удовлетворяет условию хотя бы одной из теорем при k, равном пороговому значению, то меру Жаккара для множеств биграмм, полученных из строк этих записей, не считается, поскольку итоговая мера будет заведомо меньше порогового значения.

Для записей из одного кластера считаются меры Жаккара для множеств биграмм, полученных из полей «Автор» и «Название». Если полученное значение выше порогового значения, то между этими записями устанавливается соответствие. Пороговое значение получено экспериментальным путем, оно составляет:

$$T = 0{,}7.$$

На заключительном этапе сравнения происходит проверка на исключения, так как разные части одной и той же статьи или книги имеют большое значение меры Жаккара. Для предотвращения таких случаев алгоритм отдельно просматривает строки на наличие вхождений, обозначающих номера частей: «[1]», «Часть первая», «Том 2», «Выпуск 6».

Для записей, между которыми установлено соответствие, запускается процесс создания обогащенной записи. Обогащение осуществляется по свойствам формата MODS. В библиографической записи в формате MODS свойства делятся на простые и составные. 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 встречается 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.

Сложность алгоритма выявления дублетных библиографических записей в худшем случае составляет $O(n^2)$, в лучшем случае $O(\log n)$, в среднем - $O(n)$.

Вышеописанный алгоритм реализован на языке Java с использованием фреймворка Spring MVC, объектно-реляционного отображения Hibernate и библиотеки XMLBeans.

На центральном сервере хранятся 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, проще заново создать обогащенную запись, нежели выявить разницу между первоначальной и измененной записью и применить эту разницу к обогащенной записи.

Процесс выявления дублетов и слияния записей исполняется на центральном сервере системы. Данные берутся из хранилища, в которое

попадают записи, собранные из различных библиотек с использованием протокола OAI-PMH. В результате работы алгоритма обогащенные записи попадают в отдельную базу данных, также располагающуюся на центральном сервере.

В третьей главе приводится описание выбранной онтологии для представления библиографических записей в LOD. Также описываются подходы, применяемые для хранения связанных данных и создания ссылок на внешние источники.

В разделе 3.1 рассматриваются основные форматы машиночитаемых записей, используемые в России – MARC21 и RUSMARC. Проводится исследование наиболее распространенных онтологий в контексте возможного их использования для конвертации библиографических записей. Показано, что онтологии MODS/RDF, ЕНИП РАН и Europeana Data Model являются избыточными для того набора данных, который присутствует в библиографических записях библиотек России. Предложена схема, состоящая из двух классов «Автор» и «Произведение», которая основана на использовании схем Dublin Core и FOAF. В разделе приводится полное описание разработанной схемы, а в приложении – XSLT-шаблон для преобразования данных из формата MODS в RDF/XML.

В разделе 3.2 описан алгоритм связывания библиографических записей с данными Библиотеки Конгресса США, Британской национальной библиотеки и DBpedia. Для каждого из трех сайтов на языке Java созданы программные модули, каждый из которых осуществляет поиск по заданному выражению и возвращает найденные результаты. Поскольку большая часть данных из Библиотеки конгресса США и Британской национальной библиотеки представлена на английском языке, а в библиотеках России преобладают данные на русском языке, то для повышения вероятности успешного поиска имена авторов и названия произведений предварительно переводятся на английский язык с помощью сервиса Google Translate. Для каждого источника в качестве результатов поиска модуль возвращает множество претендентов на установление связей owl:sameAs и rdfs:seeAlso.

Для отсечения заведомо неподходящих кандидатов используется метод биграмм. Исходное поисковое словосочетание и полученный результат проходят предварительную обработку, в результате 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.

Если полученное значение меры оказывается выше порогового значения

$$T_1 = 0{,}95,$$

то в RDF/XML файл добавляется RDF-триплет со связью owl:sameAs между объектами. Если полученное значение меры оказывается меньше порогового значения T_1 и выше порогового значения

$$T_2 = 0{,}7,$$

то в RDF/XML файл добавляется RDF-триплет со связью rdfs:seeAlso между объектами. Пороговые значения T_1 и T_2 установлены экспериментальным путем.

Среди обрабатываемого массива библиографических записей количество уникальных значений «Автор» и «Название» меньше, чем количество записей. Для того, чтобы не осуществлять повторные запросы к сервису Google Translate и сайтам библиотек, алгоритм поиска подобных записей сохраняет в базу данных значение исходного поля, результаты своей работы, а также время, когда данная проверка была сделана, осуществляя тем самым 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.

В разделе 3.3 производится анализ существующих RDF-хранилищ для наиболее оптимального решения предоставления доступа к RDF-триплетам с использованием протоколов SPARQL и HTTP. Рассмотрены системы 4store, Sesame и интегрированная среда Jena. В результате сравнения этих систем сделан выбор в пользу Jena.

Для загрузки RDF-триплетов в хранилище TDB, входящее в состав Jena, на языке Java написан специальный модуль. Данный модуль использует Application Programming Interface (API) для языка Java, имеющийся в Jena.

Доступ к данным по протоколу SPARQL реализован с помощью сервера Fuseki, входящего в состав Jena. Для поддержки протокола HTTP развернут и настроен веб-сервер с использованием Play Framework. Для интеграции с Fuseki используется механизм SOH – SPARQL Over HTTP, который представляет собой набор специализированных утилит – скриптов, входящих в состав Jena. В качестве результата обращения к веб-серверу по URI является RDF-файл, сериализованный с использованием Turtle. На рисунке 1 показана архитектура системы интеграции данных.

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

Раздел 4.1 посвящен Национальной электронной библиотеке. Основной целью создания НЭБ является обеспечение свободного, равного и всеобщего доступа 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.

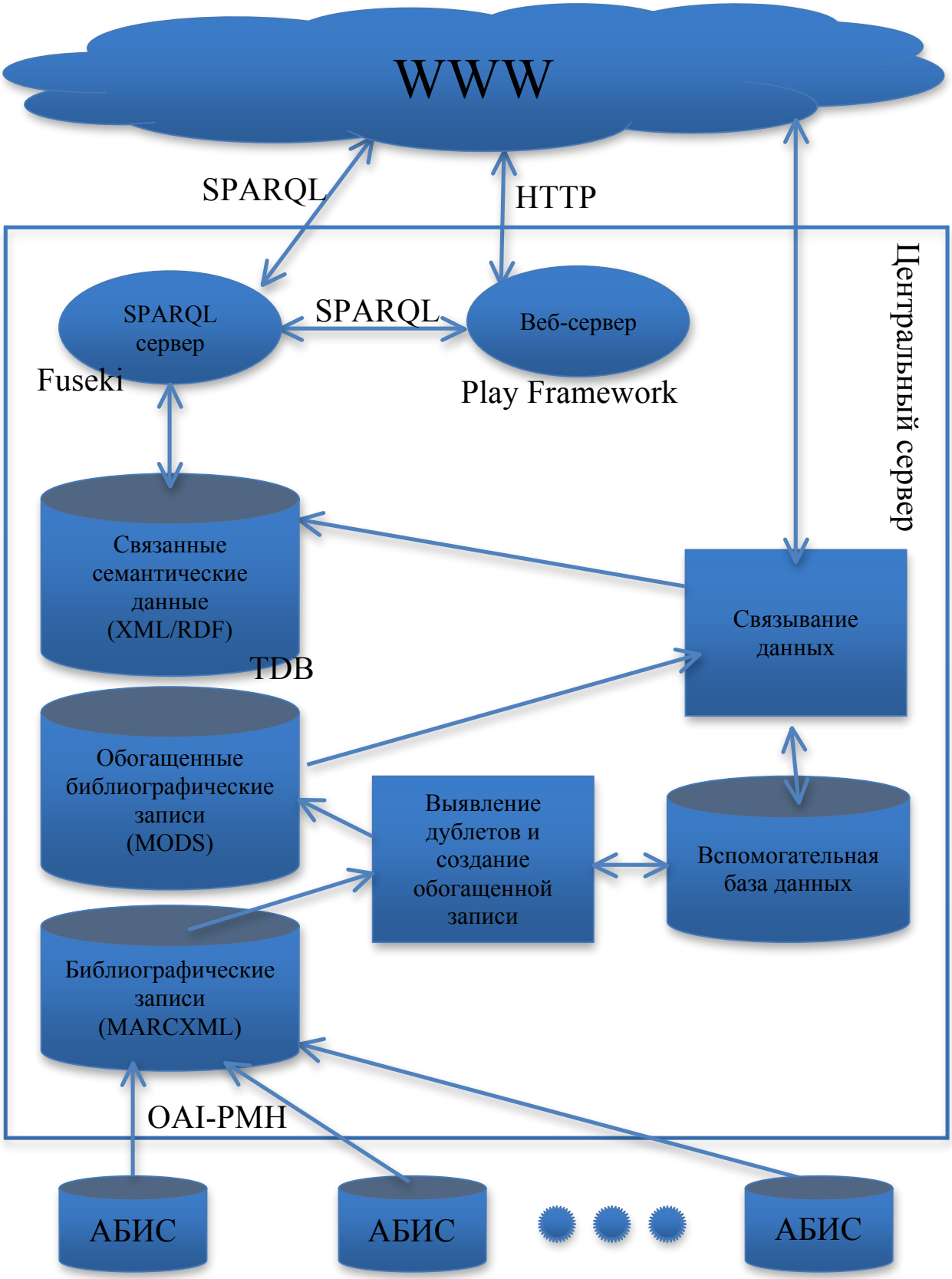

Рисунок 1: Архитектура системы интеграции данных

Основной инфраструктурной особенностью функционирования НЭБ является одновременное наличие центрального портала и распределенного характера процессов создания, каталогизации, подключения, использования и хранения ресурсов среди участников НЭБ.

В разделе подробно рассматривается строение центрального портала НЭБ, состоящего из набора подсистем, реализующих различный функционал.

В разделе 4.2 приведены результаты работы алгоритма выявления дублетных библиографических записей на электронном каталоге НЭБ, содержащем 21 313 009 записей. Для обработки такого объема данных потребовалось 90 часов. Вся информация, хранящаяся в базе данных, занимает 60 гигабайт, для ускорения работы алгоритма она полностью была подгружена в оперативную память. 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.

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 1: Количество записей с определенным количеством дублетов

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.

В разделе 4.3 приведены результаты работы алгоритма поиска и связывания с уже опубликованными данными в LOD. 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%.

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 2.

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

Таблица 2: Среднее количество RDF-триплетов связи в одном RDF/XML-файле

На рисунке 2 показан результат обращения к веб-серверу по URI – RDF-файл, сериализованный с использованием Turtle.

В заключении формулируются основные результаты и выводы диссертационной работы.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

В диссертационной работе выполнен анализ существующих и перспективных подходов, технологий, онтологий, алгоритмов и систем, применяемых для консолидации большого 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.

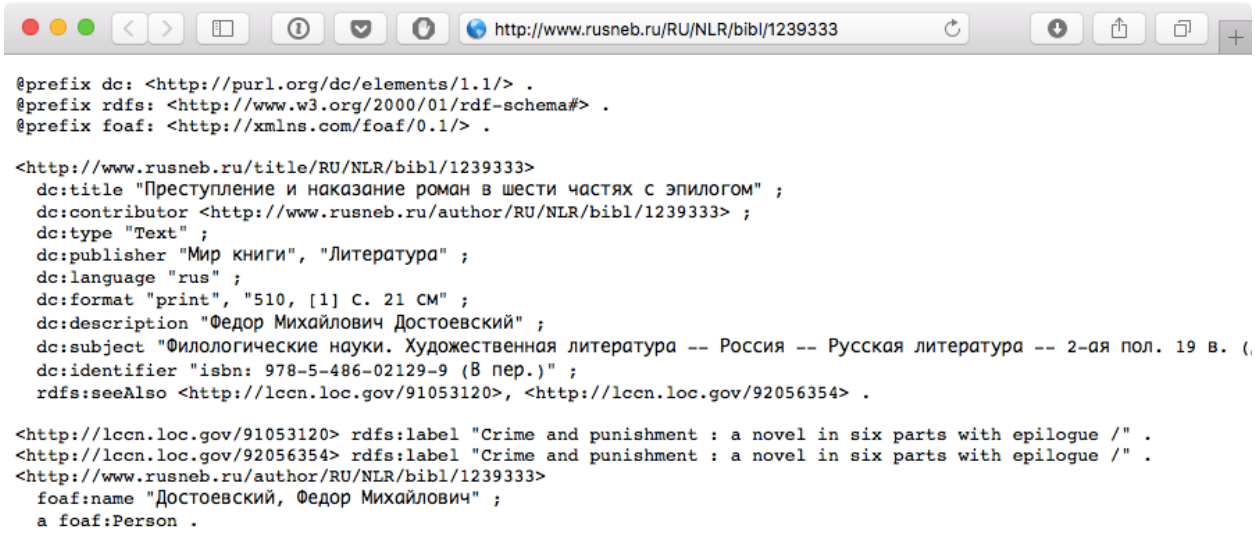

Рисунок 2: Результат обращения к веб-серверу

Основными научными и практическими результатами диссертационной работы являются:

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

Основные результаты диссертационной работы отражены в следующих опубликованных работах:

1. **Шорин, О.Н. Публикация библиографических данных в открытое семантическое пространство / О. Н. Шорин // Информационные ресурсы России.— 2016.— №4.— С. 19-23.**
2. **Шорин, О.Н. Интеграция библиографических записей / О. Н. Шорин // Информационные ресурсы России.— 2016.— №5.— С. 14-18.**
3. 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.
4. 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.
5. Шорин, О. Н. Национальная электронная библиотека: технологии реализации / О. Н. Шорин // Национальная библиотека.— 2015.— №3(06).— С. 60-65.
6. Шорин, О. Н. Проблемы семантической интеграции библиотечных данных / В. А. Серебряков, О. Н. Шорин // Библиотековедение.— 2014.— №5.— С. 41-47.
7. Шорин, О. Н. Сбор, обработка и хранение библиографических записей с использованием технологий семантической паутины / О. Н. Шорин // Библиотековедение.— 2015.— №2.— С. 37-42.
8. 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.
9. Шорин, О. Н. Современные тенденции поиска и анализа информации / О. Н. Шорин // Национальная библиотека.— 2015.— №1(04).— С. 34-39.
10. Шорин, О. Н. Стратегические IT инновации в Российской национальной библиотеке / А. В. Лихоманов, О. Н. Шорин // Медиатека и мир.— 2014.— №3.— С. 4-8.