На правах рукописи

СУЛЕЙМАНОВ РУСЛАН СУЛЕЙМАНОВИЧ

ИНТЕГРАЦИЯ ЦИФРОВЫХ ИНФОРМАЦИОННЫХ РЕСУРСОВ В ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ

Специальность 05.25.05 – Информационные системы и процессы

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Москва
2021

Диссертация выполнена в Федеральном государственном бюджетном образовательном учреждении высшего образования «Московский государственный институт культуры», на кафедре библиотечно-информационных наук

Научный руководитель:	**Гончаров Михаил Владимирович,** кандидат технических наук, доцент, руководитель группы перспективных исследований и аналитического прогнозирования ФГБУ «Государственная публичная научно-техническая библиотека».
Официальные оппоненты:	**Майстрович Татьяна Викторовна**, доктор педагогических наук, доцент, Ведущий научный сотрудник Фундаментальной библиотеки ФГБУ науки «Институт общественной информации по общественным наукам Российской академии наук».
	Соколинский Кирилл Евгеньевич кандидат технических наук, Начальник отдела информационных технологий, ГОУ ВПО Санкт-Петербургский государственный университет телекоммуникаций им. проф. М.А. Бонч-Бруевича.
Ведущая организация:	Федеральный исследовательский центр “Информатика и управление” Российской академии наук.

Защита диссертации состоится «19» марта 2021г. в 11 часов 00 минут на заседании совета по защите диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук Д 210.010.0, созданного на базе Федерального государственного бюджетном образовательного учреждения высшего образования «Московский государственный институт культуры», по адресу: 141406, Московская обл., г. Химки-6, ул. Библиотечная, д.7, корп. 2, зал защиты диссертаций (218 ауд.).

С диссертацией можно ознакомиться в Информационно – библиотечном центре и на сайте Московского государственного института культуры: http://nauka.mgik.org/

Автореферат разослан «__» _____ 2021 г.

Ученый секретарь диссертационного совета,
кандидат педагогических наук, доцент Т.Я. Кузнецова

I. Общая характеристика работы

Актуальность темы исследования. Развитие науки и образования в XXI веке невозможно без развития информационных технологий, в частности сегодня, в эпоху нарастающей цифровизации, необходим принципиально новый подход к разработке информационного обеспечения образования и науки. Быстрое развитие информационных технологий дало возможность обеспечения требуемой генерации и распространения научной и образовательной информации. Традиционно функцию накопления, использования и передачи знаний исполняли библиотеки: научные, общедоступные, специализированные. Международная федерация библиотечных ассоциаций и учреждений сообщает, что в мире сегодня существует более 569,6 тысяч традиционных библиотек, в том числе более 100 тысяч только в Российской Федерации. Отметим, что с конца 1990-х годов наряду с традиционными в практику информационного обслуживания пользователей начали входить электронные библиотеки. Одним из способов формирования и передачи знаний, обеспечивающих удобство и простоту получения информации, сегодня являются электронные библиотеки, число которых растет как в России, так и во всем мире.

На данный момент в стране создаются и используются электронные библиотеки в разных сферах деятельности, и большую роль в этом играют библиотеки и информационные центры страны, научно-исследовательские институты и образовательные учреждения. Нельзя не отметить две наиболее крупные и известные электронные библиотеки национального масштаба: «Президентскую библиотеку имени Б. Н. Ельцина» (Санкт-Петербург) и «Национальную

электронную библиотеку», оператором которой является Российская государственная библиотека. Тем не менее, несмотря на богатство электронным контентом этих двух национальных систем и существующих электронных библиотек в библиотеках, институтах и вузах страны, потребности науки, образования и культуры существенно шире. На данный момент имеется потребность, скорее даже уже требование, ученых, специалистов и обучающихся в развитии информационного обеспечения, прежде всего в расширении доступа к большому числу универсальных и профильных электронных ресурсов. Современная наука требует большого охвата разных отраслей с точки зрения создания цифровых коллекций и отдельных универсальных или проблемно-ориентированных электронных библиотек. С учетом имеющихся современных средств и инструментария необходимо учитывать не только потребности ученых, исследователей, преподавателей и студентов, но и возможности современной научной коммуникации, позволяющей оптимизировать создание, эффективное распространение и использования электронных коллекций (библиотек).

Одной из основополагающих проблем при создании электронной библиотеки является интеграция данных, так как их объем постоянно увеличивается, что приводит к тому, что становится все сложнее их интегрировать с учетом не только объема, но и форматов представления и, главным образом, обеспечения необходимой релевантности. Во-первых, требуется обеспечить непрерывный и удобный доступ для получения контента. Во-вторых, и это главное, необходимо извлечь метаданные, которые содержатся внутри документа и однозначно определяют его. При этом если сами исходные документы хранятся в одинаковых форматах, то в таком случае можно разработать правила и соответствующее программное обеспечение, позволяющие анализировать эти данные, либо

применить готовые созданные парсеры данных. Однако в случае если данные хранятся в разных форматах, не обойтись без написания собственного парсера. В качестве альтернативного метода может выступать создание универсального конструктора правил интеграции полей. Сам конструктор может быть реализован в виде веб-интерфейса, позволяющего извлекать необходимые целевые поля из документов или страниц внешней электронной библиотеки. Эта задача актуальна именно сегодня, так как позволяет повысить эффективность интеграции данных и обеспечить работу с любыми типами источников.

Одним из основных критериев удобства пользования электронной библиотекой является возможность быстро найти искомый документ, что обеспечивается поиском по метаданным. Тем не менее иногда необходимые документы публикуются в разных коллекциях, в том числе в виде файлов на диске, не сопровождаемых достаточным набором метаданных, необходимым для релевантного поиска, что создает пользователю большие проблемы при поиске документа. Однако при этом само содержимое документов может включать нужные данные: название документа, фамилию автора, информацию об издательстве и так далее. В данном исследовании рассматривается способ извлечения метаданных из полных текстов документов для повышения уровня их идентификации в электронной библиотеке.

Таким образом, актуальность данного исследования обоснована необходимостью проектирования электронных библиотек с учетом разнородности и распределенности представления данных и обеспечения требований, предъявляемых к ресурсам Интернета: прежде всего быстроты отклика на запрос, интуитивно понятного и удобного в использовании интерфейса, а также возможности интеграции ресурсов из максимального количества источников на основании использования метаданных.

Степень научной разработанности разных аспектов темы исследования достаточно высока. В отечественной и зарубежной библиотечно-информационной науке в последние годы подготовлено немало статей и обзоров по различным вопросам, вошедшим в рамки изучаемой в исследовании проблемы.

В своей работе автор опирался на методологии проектирования электронных библиотек, введенные Антопольским А.Б., Земсковым А.И., Шрайбергом Я.Л. Вопросы, связанные с организацией работы поисковых библиотечных систем были затронуты в работах Каленова Н.Е., Колосова К.А., Соколинского К.Е., Сотникова А.Н. Проблемы интеграции информации из распределенных источников описаны в трудах Погорелко К.П., Рябова В.И., Серебрякова В.А., Соболевской И.Н. Проблемы интеграции существующих библиотечных ресурсов в единую базу данных решаются в таких проектах как “Научное наследие России”.

Научные труды многих известных ученых позволили определить цель исследования, однако анализ имеющейся литературы показал недостаточную степень изученности проблемы интеграции информации из распределенных источников с применением методики извлечения метаданных из полнотекстовых электронных документов, что является одним из наиболее значимых аргументов для подготовки настоящего диссертационного исследования.

Целью настоящего исследования является улучшение качества интеграции цифровых информационных ресурсов из разных источников с помощью модели и методики, учитывающих разные структуры данных.

Задачи исследования.

1. Провести анализ существующих способов и механизмов интеграции данных в электронных библиотеках.

2. Разработать модель и спроектировать эффективный конструктор правил интеграции информации из

распределённых источников (базы данных, веб-сайты и полнотекстовые документы в формате PDF).

3. Разработать методику извлечения метаданных из полных текстов оцифрованных документов, что позволит повысить полноту предоставленных метаданных текстов на естественном языке.

В качестве **теоретической и методологической основы диссертации** выступают исследования и разработки отечественных и иностранных ученых в области построения баз данных, интеграции материалов библиотек, извлечения метаданных.

При работе над диссертацией автором были использованы труды российских и зарубежных ученых Антопольского А.Б., Вислого А.И., Гончарова М.В., Земскова А.И., Калёнова Н.Е., Колосова К.А., Лопатиной Н.В., Лютецкого В.М., Мазурицкого А.М., Соколинского К.Е., Сотникова А.Н., Тютюнника В.М., Цветковой В.А., Шрайберга Я.Л., Tillett B.B. и других.

Научная новизна состоит в обосновании и разработке новой экспериментальной методики интеграции цифровых данных из разнородных и распределённых источников для электронных библиотек. Методика позволяет выявить качественно новые закономерности представления метаданных в цифровых документах, являющихся единицей записи данных из коллекций в электронных библиотеках:

1. Разработана и обоснована модель конструктора правил интеграции информации из распределенных источников для электронных библиотек, позволяющая упростить процесс наполнения базы данных электронных документов и доказавшая перспективность для использования в построении электронных библиотек.

2. Разработана и обоснована методика извлечения метаданных, в том числе новые грамматики и словари на основе естественного языка, используемая для анализа

полнотекстовых оцифрованных документов, а также программная реализация механизма извлечения метаданных из полнотекстовых документов.

Теоретическая значимость проблемы интеграции информации из распределенных источников, возникающие в основном из-за разных форматов хранения метаданных.

Создана и обоснована модель «Конструктора правил интеграции электронных документов из распределенных источников для электронных библиотек». Теоретическая значимость данной модели заключается в расширении представлений о механизмах формирования электронных библиотек, которая, в том числе, раскрывает особенности построения справочно-поискового аппарата электронных библиотек.

Применительно к проблематике диссертации результативно использована методика извлечения метаданных, применяемая для анализа полнотекстовых оцифрованных документов.

Практическая значимость.

Теоретические и экспериментальные результаты, полученные в ходе диссертационного исследования, прошли апробацию и были внедрены в Московском педагогическом государственном университете. Разработанные методики используются в управлении фондом электронной библиотеки Московского педагогического государственного университета. Отдельные модули автоматизированной системы управления электронной библиотекой и модуля интеграции данных используются в управлении библиотечным фондом Московского городского педагогического университета.

В открытый репозиторий по лицензии GNU General Public License (универсальная общественная лицензия GNU) выложен исходный код конструктора правил интеграции информации из распределенных источников, который

позволяет автоматизировать сбор и обработку данных и метаданных оцифрованных печатных документов, в том числе книг.

Разработанный конструктор позволил объединить имеющиеся оцифрованные материалы для электронной библиотеки Московского педагогического государственного университета и автоматизировать управление фондом библиотеки Московского городского педагогического университета в части наполнения электронной библиотеки метаданными.

Результаты диссертационного исследования были использованы в управлении фондом библиотеки Московского педагогического государственного университета, что подтверждается наличием справки о внедрении.

По результатам диссертационного исследования были зарегистрированы две программы для ЭВМ: № 2012619529 - «Система управления контентом электронной библиотеки», дата регистрации 22.10.2012 (совместно с Шабановым Б.М., вклад автора диссертации - постановка задачи); № 2019661660 - «Конструктор правил интеграции данных для электронных библиотек», дата регистрации 05.09.2019 (без соавторов).

Достоверность полученных научных результатов подтверждена результатами практических применений, положительными результатами их обсуждения на научных конференциях.

Апробация работы.

Основные положения работы докладывались на XI научно-практической конференции «Современные информационные технологии в управлении и образовании» (Москва, 2012); XVII научно-практическом семинаре «Информационное обеспечение науки: новые технологии» (Таруса, 2013); III международной научно-практической конференции Innovative Information Technologies (Прага,

2014); на Московском международном салоне образования в 2018 и 2019 годах.

Личный вклад.

Автором самостоятельно поставлены цель и задачи работы, разработана структура базы данных электронной библиотеки, позволяющая интегрировать информацию из разных источников, разработан конструктор полей интеграции данных, разработан метод извлечения метаданных из полнотекстовых документов, разработана программа эксперимента, проведен анализ результатов эксперимента и выявлены основные закономерности извлечения метаданных.

Результаты научного исследования отражены в семи публикациях, большая часть публикаций сделана лично соискателем, в том числе две статьи в журналах, рекомендуемых ВАК для публикации результатов диссертаций на соискание ученой степени кандидата технических наук по специальности 05.25.05.

Методология и методы исследования.

В работе использованы структурного анализа, системного анализа теории проектирования баз данных, теории объектно-ориентированного программирования, теории анализа текстов на естественном языке.

Программное обеспечение для прогностической части работы реализовано средствами языка PHP в связке с СУБД MySQL, поисковой машины Sphinx и Яндекс «Томита-парсер».

Объектом исследования является структура информационного массива, позволяющего интегрировать данные из распределенных источников.

Предметом исследования является методика интеграции данных в электронные библиотеки.

Основные положения, выносимые на защиту:

1. Анализ имеющихся способов интеграции информации из распределённых источников выявил проблемы,

возникающие в основном из-за разных форматов хранения метаданных в электронных библиотеках.

2. Для решения проблем интеграции информации из распределённых источников в разных форматах хранения метаданных разработана и обоснована модель конструктора правил интеграции информации из распределённых источников для электронных библиотек, позволяющая упростить процесс наполнения базы данных электронных документов. Модель была апробирована и доказала перспективность для использования в построении электронных библиотек.

3. Для повышения полноты предоставленных метаданных в электронных библиотеках разработана и обоснована методика извлечения метаданных, в том числе новые грамматики и словари на основе естественного языка, используемая для анализа полнотекстовых оцифрованных документов, а также программная реализация механизма извлечения метаданных из полнотекстовых документов.

Структура диссертации.

Цели и задачи диссертации обусловили ее структуру. Работа состоит из введения, двух глав, основных выводов по каждой главе, заключения, списка использованных источников и приложений. Диссертация содержит 131 страницу машинописного текста, 12 рисунков и 3 таблицы. Библиография включает 96 наименований.

II. Основное содержание диссертации

Во Введении обосновывается актуальность выбранной темы диссертационного исследования, характеризуется **степень ее разработанности**, определяются цели и задачи, осуществляется выбор предмета и объекта исследования, определяются методологические основания исследования,

теоретическая и практическая значимость результатов исследования.

В **первой главе** автор привел базовые определения основных терминов и описал краткую историю появления и развития электронных библиотек. Были рассмотрены методы хранения данных в электронных библиотеках, описаны основные характеристики, процессы и различия электронных библиотек от традиционных. Приведены изначальные принципы при проектировании автоматизированной библиотечно-информационной среды. Даны определения метаданных, стандартов и форматов их хранения, а также типизация информационных объектов.

Автор производит оценку объемов информации в интернете, благодаря мониторингу объемов трафика. В главе подробно рассматриваются способы получения требуемой информации пользователями библиотек (в том числе электронных), рассматриваются преимущества и недостатки каждого способа.

В главе подробно рассматривается история появления и характеристики информационных массивов. Электронные библиотеки являются одним из подвидов информационных массивов и обладают своими собственными особенностями и характеристиками. Как и к традиционным библиотекам, к электронным библиотекам может быть применена разная классификация, однако классификация электронных библиотек более разнообразна за счет различия в возможностях преподнесения информации.

Создание электронной библиотеки требует определения общесистемных требований и правил к разработке. Автор описывает особенности проектирования электронных библиотек как одного из видов информационных систем и связывает с ними основные принципы и технологии Семантической паутины.

Как бы хорошо не была спроектирована электронная библиотека – наиболее важным аспектом является ее контент. Сам контент электронной библиотеки можно разделить на две части: данные (материалы) и их метаданные. Метаданные описывают свойства того или иного материала в библиотеке, показывая различную информацию о нем – сведения о наименовании, авторе, издательстве и любые другие. Сами материалы – это непосредственно те ресурсы, которые нужны пользователям, то есть книги, статьи и любой другой контент. Метаданные предназначены для упрощения поиска, каталогизации и разделения ресурсов друг от друга. Автор описывает существующие и общепринятые словари основных понятий для описания и унификации метаданных на примере Dublin Core, а также форматы хранения MARC, UNIMARC, RUSMARC и дает оценку сложности интеграции данных из различных электронных библиотек, хранящих данные в различных форматах.

Во **второй главе** автором рассмотрена концептуальная схема электронной библиотеки FRBR, описаны варианты связей в базе данных электронной библиотеки, предложена структурная схема электронной библиотеки, позволяющая обеспечить соответствие библиотеки предъявляемым требования, а также рассмотрена проблема интеграции данных из различных источников.

Автором предлагается общая схема основных таблиц базы данных электронной библиотеки, в которой используются те же элементы, что и в модели FRBR, но структура более детализирована и содержит концепцию объединения отдельных элементов в упорядоченную базу данных.

В предложенной структуре базы данных электронной библиотеки между всеми таблицами существует связь «многие ко многим» за счет использования промежуточных таблиц. Такая структура позволяет обеспечивать максимальную

навигацию в рамках библиотеки, извлекать любые данные из таблиц, а также формировать сложные запросы.

Предложенная структура также подразумевает наполнение библиотеки как собственными ресурсами, так и интеграцию с другими базами данных, что позволяет обеспечить максимальную наполняемость библиотеки, и, как следствие, ее универсальность и разнообразность. Существует ряд способов интеграции данных в тех случаях, когда форматы хранения совпадают или же достаточно описаны для создания специализированного конвертера, таких как:

1. интеграция на уровне полей базы данных;
2. использование API;
3. получение метаданных по коду ISBN;
4. синтаксический разбор HTML страниц исходной электронной библиотеки;
5. извлечение метаданных непосредственно из текстов материалов исходной электронной библиотеки.

Автор рассматривает особенности, преимущества и сложности в использовании каждого способа.

Предложенная схема структуры электронной библиотеки обеспечивает эффективную навигацию, минимальное время ответа на запрос пользователя и возможность интеграции с материалами других библиотек способами, описанными выше. Однако при необходимости интеграции материалов более чем из одного источника процесс является трудоемким. Например, для синтаксического разбора HTML страниц исходной электронной библиотеки требуется создание парсера (синтаксического анализатора) HTML-кода для каждой исходной библиотеки, а также робота для автоматизированного обхода страниц целевой библиотеки.

На программную реализацию предложенной автором схемы электронной библиотеки было получено свидетельство о государственной регистрации программы для ЭВМ №2102619529.

Для решения задачи интеграции материалов более чем из одного источника автором был спроектирован и реализован конструктор полей интеграции данных.

Часть главы посвящена выбору языка и программных средств для реализации проекта. Представляется сравнительный анализ платформ для автоматизации библиотек и приводятся системные требования к ним.

Далее производится сравнение полей и построение связей для обеспечения возможностей интеграции данных из различных форматов, на основе чего создается обобщенная модель.

После построения модели предлагается метод реализации алгоритма построения конструктора полей интеграции данных из различных источников.

Схема взаимодействия с конструктором может представлять собой пошаговый механизм создания правил для интерпретации исходных полей в требуемые. Конструктор может быть реализован как WEB-приложение, запускаемое в браузере и работающее с базой данных через ассинхронные запросы методом Ajax. Перед началом работы требуется произвести базовую настройку, которая заключается в выборе типа обрабатываемых данных (поля базы данных, данных в форматах JSON или XML, либо отдельные элементы из модели DOM (Document Object Model — «объектная модель документа») HTML страницы). Разбор модели DOM HTML-документов является наиболее трудоемким из-за того, что в отличие от остальных типов форматов его содержимое не является формализованным, так как модель DOM не накладывает ограничений на структуру документа. Любой документ известной структуры с помощью DOM может быть представлен в виде дерева узлов, каждый узел которого представляет собой элемент, атрибут, текстовый, графический или любой другой объект. Узлы связаны между собой отношениями "родительский-дочерний". Для выбора

отдельных частей документа внутри DOM могут быть использованы как простые CSS-селекторы, такие как элементы, ID, Class, так и вложенные, например, селектор потомков (контекстный селектор), селектор дочерних элементов или селектор сестринских элементов.

После настройки конструктора и назначения правил интеграции необходимо выбрать источник входных данных. В качестве источника можно указать отдельный файл (или список файлов), таблицы базы данных с указанием доступа или ссылки на API в форматах XML/JSON. В случае выбора доступа по ссылкам имеется настройка ввода правильного и неправильного ответа, что необходимо в случае, если получаемая информация разбивается на страницы или предусмотрены коды ответов в зависимости от статуса запроса.

Результаты работы конструктора можно получить через RESTful API (Representational State Transfer — «передача репрезентативного состояния»). REST - метод взаимодействия компонентов распределённого приложения в сети Интернет, при котором вызов удаленной процедуры представляет собой обычный HTTP-запрос (обычно GET или POST, такой запрос называют REST-запрос), а необходимые данные передаются в качестве параметров запроса.

Предлагаемый конструктор имеет богатый функционал для работы с мета-данными различных типов. Однако, в случае, если нужно интегрировать электронные версии документов, не сопровождаемые мета-данными из обычной файловой системы, то его использование приведет к нулевым результатам. Для решения данной задачи автор предлагает использовать средства для извлечения мета-данных из полных текстов оцифрованных версий книжных материалов при помощи Яндекс “Томита-парсера”, предназначенного для разбора по недетерменированным и неоднозначным грамматикам, и специального конвертера. В качестве входных

материалов можно использовать файлы в формате Adobe PDF с текстовой подложкой. Выбор формата хранения материалов обусловлен широкой распространенностью документов данного типа, а также наличием специальных утилит для экспортирования “чистого” текста из них, например, «pdftotext» или «pdf2rtf».

Томита-парсер — это инструмент для извлечения структурированных данных (фактов) из текста на естественном языке. Извлечение фактов происходит при помощи контекстно-свободных грамматик и словарей ключевых слов.

Для решения задачи извлечения мета-данных на естественном языке при помощи Томита-парсера требуется создать следующие компоненты:

1. КС-грамматики (набор правил, описывающих синтаксическую структуру извлекаемых цепочек слов);
2. Газзетиры (словари с ключевыми словами для грамматик);
3. файлы, описывающие факты (регулирует механизм преобразования грамматик в конкретные факты).

Для разбора полных текстов предлагается создать программу (конвертер), принимающую на входе путь к каталогу с файлами в формате PDF и следующую алгоритму:

1. Обход исходного каталога, построение индекса файлов;
2. Извлечение полных текстов из всех найденных файлов;
3. Извлечение фактов из полных текстов;
4. Программная обработка полученных мета-данных (очистка от лишних символов, изменение словоформ, объединение различных фактов.);
5. Создание выходного файла, включающего обработанные данные;
6. Импорт выходного файла в созданный конструктор правил интеграции.

В рамках исследования были разработаны словари и грамматики для извлечения следующих мета-данных, являющихся обязательной информацией об издании и требующихся при публикации в электронной библиотеке для обеспечения каталогизации и доступности материалов:

— название материала;
— сведения об авторах;
— код ISBN (уникальный номер книжного издания);
— год публикации;
— место публикации;
— сведения об издателе;
— коды рубрикаторов (УДК, ББК, ГРНТИ).

Для подтверждения корректности извлечения мета-данных из полных текстов электронных версий печатных материалов был проведен ряд экспериментов. В качестве тестовой площадки для проведения экспериментов был использован набор данных, содержащий 10 000 русскоязычных книг из базы данных электронной библиотеки «Научное наследие России». Для всех материалов, на которых производилось тестирование, были доступны мета-данные, с которыми было проведено сравнение извлеченных данных методом сравнения полей.

В результате проведения экспериментов были получены следующие результаты, представленные в табл.1.

Таблица 1. Корректность извлечения мета-данных из тестовой выборки материалов

Поле	**Извлечено верно (%)**	**Извлечено неверно (%)**	**Требуется уточнение (%)[1]**
Наименование материала	76	21	3
Сведения об авторах	91	7	2
Код ISBN	98	0	2
Год публикации	89	10	1
Место публикации	84	12	4
Сведения об издателе	79	14	7
Коды рубрикаторов	90	1	9
Результаты в среднем	**86,7**	**9,3**	**4**

Средний показатель корректно извлеченных мета-данных составил 86,7%, еще 4% извлеченных фактов поддаются последующей корректировке и могут быть использованы после ее проведения. При этом наибольшие проблемы наблюдаются с извлечением наименований материалов, которые не имеют четко утвержденной структуры, могут содержать любое количество символов и знаков препинания.

[1]Выявлены ошибки при оптическом распознавании текста (OCR), данные извлечены не полностью, либо извлечена лишняя информация.

Процент успешного извлечения метаданных из полных текстов можно увеличить благодаря улучшению качества оптического распознавая печатных материалов, а также улучшению КС-грамматик и газзетиров.

Таким образом, возможна интеграция даже тех материалов, которые не сопровождаются мета-данными.

Благодаря разработанному алгоритму извлеченные факты могут быть импортированы в базу данных после создания соответствующих правил и настройки полей при помощи визуального редактора.

Таким образом можно интегрировать материалы только благодаря их полными текстам, даже если они не сопровождаются никакими внешними мета-данными, что может помочь значительно улучшить показатели успешной интеграции материалов исходной библиотеки, либо снизить временные затраты при работе администратора или редактора электронной библиотеки.

Спроектированная модель конструктора полей интеграции данных позволяет извлекать целевые поля из данных любой электронной библиотеки, доступ к которой осуществляется через Интернет, так как предназначен для работы с любыми типами источников. Успешное извлечение данных обеспечивается перечислением требуемых внешних источников данных и построению правил (синтаксис правил зависит от формата источника), с помощью чего присваивается соответствие исходных и целевых полей в базе данных электронной библиотеки, что позволяет избавиться от необходимости ручного копирования мета-данных материалов внешней библиотеки.

Далее описывается механизм объединения всех созданных элементов АСУ электронной библиотекой.

III. Заключение

В заключении подводятся итоги диссертационного исследования, излагаются его основные выводы и обобщающие результаты.

В Результате проведенного диссертационного исследования:

1. Проведен анализ существующих способов и механизмов интеграции данных в электронных библиотеках;

2. Разработана модель и спроектирован эффективный конструктор правил интеграции информации из распределенных источников (базы данных, веб-сайты и полнотекстовые документы в формате PDF);

3. Разработана методика извлечения метаданных из полных текстов оцифрованных печатных документов, что позволяет повысить полноту предоставленных метаданных текстов на естественном языке.

Все поставленные задачи и цель диссертационного исследования - улучшение качества интеграции цифровых информационных ресурсов из различных источников с учетом различных структур данных – достигнута.

Список работ, опубликованных автором по теме диссертации

Публикации в изданиях, рекомендованных ВАК России*:*

1. Сулейманов Р. С. Сбор библиотечной информации из распределенных электронных источников при помощи конструктора правил интеграции данных / Р.С. Сулейманов // Информационные ресурсы России. - 2016. - № 6. - С. 23-26.
2. Сулейманов Р. С. Современные подходы к интеграции данных в электронных библиотеках / Р.С. Сулейманов // Информационные ресурсы России. - 2019. - № 6. - С. 13-16.

Публикации в других изданиях*:*

1. Сулейманов Р. С. Извлечение метаданных из полнотекстовых электронных русскоязычных изданий при помощи томита-парсера / Р.С. Сулейманов // Программные продукты и системы. - 2016. - № 4. - С. 58-62.
2. Сулейманов Р. С. Социальная сеть РАН - единое информационное пространство для ученых / Р.С. Сулейманов // Программные продукты и системы. - 2012. - № 4. - С. 46-49.
3. Каракозов С. Д. Техническая политика и этапы развития цифровой образовательной среды МПГУ / С. Д. Каракозов, Р. С. Сулейманов, А. Ю. Уваров // Наука и школа. - 2015. - № 1. - С. 17-27.
4. Savin G. I. Comparative analysis of solutions for full-text search in digital libraries / G. I. Savin, A. N Sotnikov, R. S. Suleymanov // Innovative information technologies : proc. of the 3-rd Intern. sci.-practical conf., Prague, 21-25 Apr.

2014. - Moscow : HSE, 2014. - Part 2 : Innovative information technologies in science. - P. 624-629.

Регистрации программного продукта:

1. Свидетельство о государственной регистрации программы для ЭВМ №2102619529. «Система управления контентом электронной библиотеки». Зарегистрировано в Реестре программ для ЭВМ от 22 октября 2012 г.
2. Свидетельство о государственной регистрации программы для ЭВМ №2019661660. "Конструктор правил интеграции данных для электронных библиотек". Зарегистрировано в Реестре программ для ЭВМ от 05 сентября 2019 г.