УДК 004

БИБЛИОТЕКА НАУЧНЫХ ПРЕДМЕТНЫХ ОБЛАСТЕЙ SCILIBRU

О. М. Атаева[1] [0000-0003-0367-5575], Н. П. Тучкова[2] [0000-0001-5357-9640],
К. Б. Теймуразов[3] [0009-0000-0876-4222], А. Абдышов[4] [0009-0000-4971-0378],
М. Г. Кобук[5] [0009-0002-9834-8218]

[1–3]*Федеральный исследовательский центр «Информатика и управление» Российской академии наук, г. Москва, Россия*
[1, 4, 5]*Московский университет имени С.Ю. Витте, г. Москва, Россия*

[1]OAtaeva@frccsc.ru, [2]NTuchkova@frccsc.ru, [3]KTeymurazov@frccsc.ru,
[4]abdysovajdin@gmail.com, [5]mikhail.kobuk@mail.ru

Аннотация

Работа посвящена проблеме интеграции данных для представления научных предметных областей на основе их семантического описания в цифровой библиотеке SciLibRu. В качестве модели данных использованы онтология и граф знаний библиотеки LibMeta. Наполнение библиотеки SciLibRu осуществляется путем добавления данных научных журналов. Показано, как реализованы этапы анализа слабоструктурированных научных публикаций для их встраивания в онтологию библиотеки. При прохождении всех этапов предобработки данных формируется датасет, который может быть использован в обучении языковых моделей для запросов в русскоязычных научных предметных областях.

Приложение работы заключается в создании рекомендательных систем для работы с научными русскоязычными журналами.

Ключевые слова*: прикладная онтология, граф знаний, источники данных,* анализ слабоструктурированных научных публикаций*.*

ВВЕДЕНИЕ

Библиотека предметных областей SciLibRu сформирована в рамках подхода, который использовался в семантической библиотеке LibMeta [1], на основе онтологического проектирования [2] и навигации по данным с помощью графа

знаний (knowledge graph, KG) [3]. LibMeta – это цифровая библиотека, данные которой связаны иерархическими и ассоциативными отношениями в соответствии с *онтологией предметной области*.

Онтология цифровой семантической библиотеки представляет собой формальное описание множества данных (типы данных, связи) предметной области (subject domain, SjD) [4]. Для построения онтологии был использован язык OWL (https://www.w3.org/2006/04/OWL_UseCases-ru.html).

Онтология на OWL представлена в виде RDF-графа (https://www.w3.org/TR/rdf12-schema/). Описание на OWL содержит *структуры* данных, а *не самих данных*. Каждый экземпляр данных – это экземпляр элемента онтологии. В библиотеке LibMeta тезаурус [5] представлен онтологией на OWL.

Процесс редактирования библиотеки LibMeta заключается в редактировании *онтологии*. В результате интеграции данных в библиотеке LibMeta накоплены описания SjD математики и смежных областей, включая описания классификаторов, научных журналов и других источников, которые семантически связаны в онтологии библиотеки SciLibRu. Развитие навигации в библиотеке с применением KG позволяет перейти к поиску, задав в поисковом запросе принадлежность данных SjD.

В настоящей работе предложено описание коллекции данных SciLibRu, которые отличаются от данных LibMeta наличием новых связей, полученных в результате метрического анализа [6, 7] KG LibMeta и достраивания онтологии, хотя эти множества, конечно, пересекаются. Описание данных SciLibRu стало необходимым исследованием, поскольку дальнейшее развитие библиотеки предполагает постановку задач о применении больших языковых моделей (Large LLanguage Models, LLM) и создании рекомендаций для извлечения знаний, накопленных в этой библиотеке. Множество данных LLM, предварительно обученной на общих данных, предполагается дополнять данными KG SciLibRu, чтобы ограничивать LLM рамками предметной области, задаваемой KG.

В работе представлены описание семантической модели данных и KG, оценка качества данных для интегрированных в библиотеку энциклопедий, пример интеграции LLM и KG для SjD обыкновенных дифференциальных уравнений

(ordinary differential equation, ODE), описана процедура гибридного алгоритма загрузки слабоструктурированных данных журналов, содержащих символьную информацию (формулы).

1. СЕМАНТИЧЕСКАЯ МОДЕЛЬ И ДАННЫЕ SCILIBRU

1.1. Семантическая модель SciLibRu

Будем использовать понятия *тезауруса* предметной области [8, 9], *онтологии* [4, 7], *KG* (*графа знаний*) [10, 11], *знаний* [12, 13]. Данные SciLibRu состоят из множества научных статей, энциклопедий, классификаторов, словарей, тезаурусов, корпусов текстов и других источников оцифрованной информации, представленного в виде онтологии SjD. *Онтология* SciLibRu реализует семантическую модель данных, в основе которой лежит тезаурус предметной области. Онтология проецируется на KG, поскольку реализуется через схему RDF. Основной контент SciLibRu составляют математические SjD, приложения и смежные области из прикладных и междисциплинарных журналов.

Знания трактуются как структурированные *данные*. Извлечение знаний состоит в получении ответа на информационный запрос. Ответ может быть (не)релевантным теме запроса и (не)пертинентным, то есть (не)удовлетворяющим информационную потребность пользователя, (не)соответствующим тезаурусу адресата [8].

Семантическая модель данных представляет собой структурированные данные, где есть объект, субъект и отношения между ними. Именно отношения составляют смысл этой тройки данных. В KG отношения – это связи графа, а объекты и субъекты – вершины графа.

Онтология LibMeta [1] – трехуровневая, содержит:

- универсальные *понятия* онтологии;
- описания *объектов* прикладной области;
- *метаданные прикладной области* как таковые.

Основными элементами онтологии на OWL являются описания классов, их свойств, отношений между классами и представителями (индивидов) классов (их свойств и отношений). Для этих описаний OWL последовательно использует бинарные отношения своего словаря, а также словарей RDF и RDFS (https://www.w3.org/TR/owl-semantics/).

SjD LibMeta, *DLib* = {*Per, Pub, Th, Enc, Jour*}:

- данные персон и их свойств, множество {*Per*};
- данные публикаций и их свойств, множество {*Pub*};
- данные тезаурусов, множество {*Th*};
- данные энциклопедий, множество {*Enc*};
- данные журналов, множество {*Jour*}.

Данные SjD SciLibRu DSci содержат множество DLib и данные множества {KG}, то есть DSci ={*Per, Pub, Th, Enc, Jour, KG*}.

На рис. 1 показано, что к трехуровневой модели онтологии LibMeta добавился уровень данных KG.

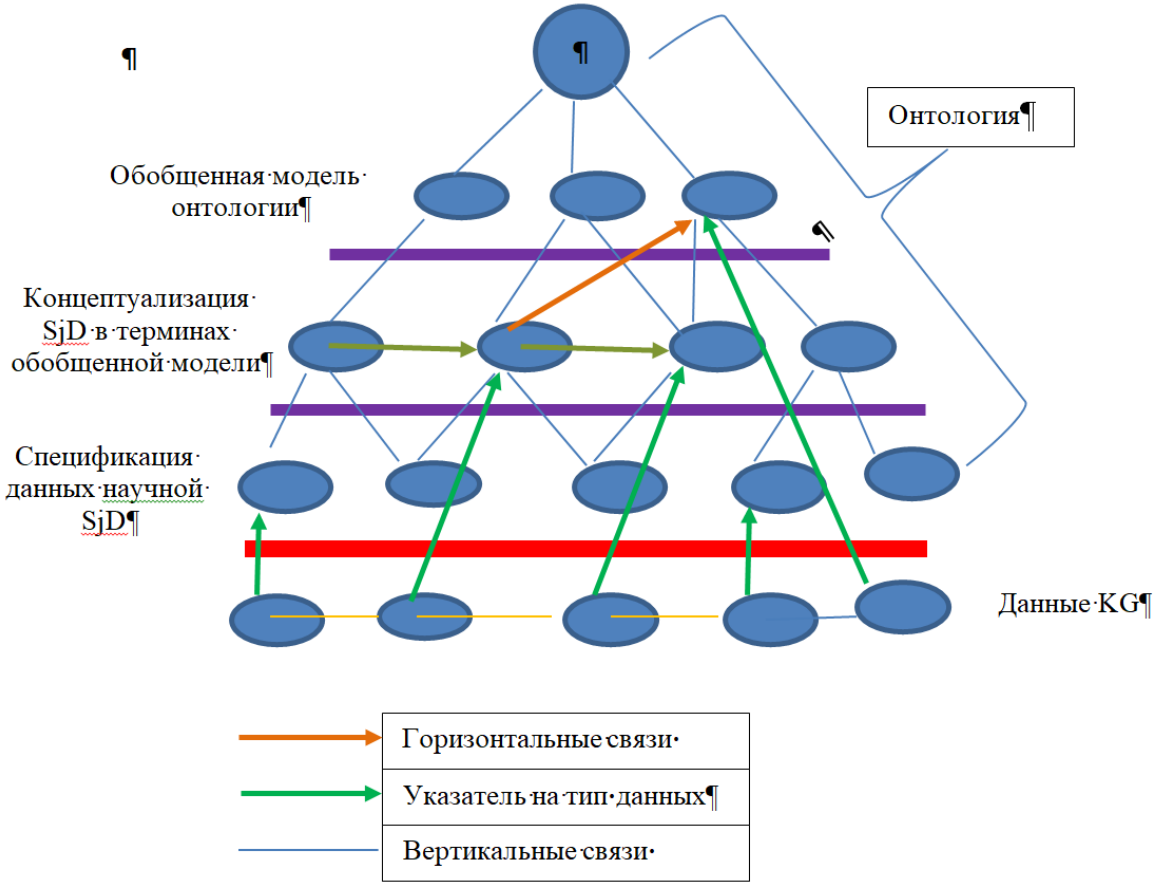

Рис. 1. Архитектура данных в *SciLibRu*.

1.2. Анализ качества KG SciLibRu

Оценка качества данных стала одним из этапов представления знаний, поскольку внедрение LLM «перекладывает» вопросы достоверности с экспертов научных областей на создателей информационных систем, которые организуют интеграцию данных и навигацию по ним. Известные недостатки применения LLM в поисковых запросах, которые приводят к «галлюцинациям», связаны как с логикой модели, так и с массивом данных (датасет), на которых проводилось обучение языковой модели [14].

Для библиотеки SciLibRu оценка данных заключается в проверке связей в онтологии и, соответственно, в KG, как в результирующем множестве концептов и связей, который будет использоваться на входе LLM [15].

Для нашего исследования были вычислены метрики KG SciLibRu для вершин классификаторов УДК (https://teacode.com/online/udc/) и ГРНТИ (https://grnti.ru/) и получены значения метрик для оценки качества графа.

Общее количество узлов, число вершин (узлов) в графе: $|V| = 10813$.

Общее количество связей, число ребер (связей) в графе: $|E| = 178728$.

Распределение узлов KG SciLibRu по типам**:**

тип 1 – KG математической энциклопедии [16], $|V|_1 = 6263$,

тип 2 – KG энциклопедии математической физики [17], $|V|_2 = 3228$.

Распределение узлов KG SciLibRu по связям показывает, сколько ассоциативных связей (ребер) каждого типа в KG:

$|E|_1 = 155448$ (связь *related*), $|E|_2 = 11805$ (связь *use*), $|E|_3 = 11475$ (связь *seealso*).

Плотность – мера *заполненности* графа связями: $\rho = \frac{|E|}{|V|(|V|-1)}$.

Для направленного графа: $|V|(|V|-1)$ – это возможное максимальное количество ребер (каждый узел может иметь ребро к каждому, кроме себя). Плотность графа KG SciLibRu: $\rho = 0.0015$.

Низкое значение плотности графа указывает на его *разряженность*, что типично для графов большого размера (граф энциклопедий) и объясняется тем, что большинство узлов связано только с небольшим подмножеством других узлов преимущественно в рамках тематических взаимосвязей каждого понятия.

В графе KG SciLibRu выявлено 1057 изолированных узлов типа *Concept*, что составляет около 10% от общего количества понятий SjD в графе. Это указы-

вает на отсутствие связей этих понятий с другими сущностями, что может свидетельствовать о недостаточности доступной информации в данной подобласти. Далее были вычислены метрики центральности**.**

Самыми влиятельными узлами оказались наиболее общие понятия, что говорит о высокой взаимосвязанности и значимости этих понятий в охватываемых SjD. Анализ центральности подтвердил значимость этих понятий, при этом были выявлены дополнительные узлы, которые, несмотря на меньшую распространенность, обладают высокой вероятностной значимостью в графе. Кроме того, в результате анализа были обнаружены узлы, замыкающиеся на себя (пример эго-графа «Функция», рис. 2), то есть их (концепты и связи) необходимо дополнительно анализировать. В поисковой выдаче, вероятно, это приведет к дублированию.

Качественный анализ графа показал, что есть связи, которые не были выявлены в процессе предобработки данных, что тоже, очевидно, влияет на результат поиска. Была установлена корреляция между эмбеддингами и классическими метриками.

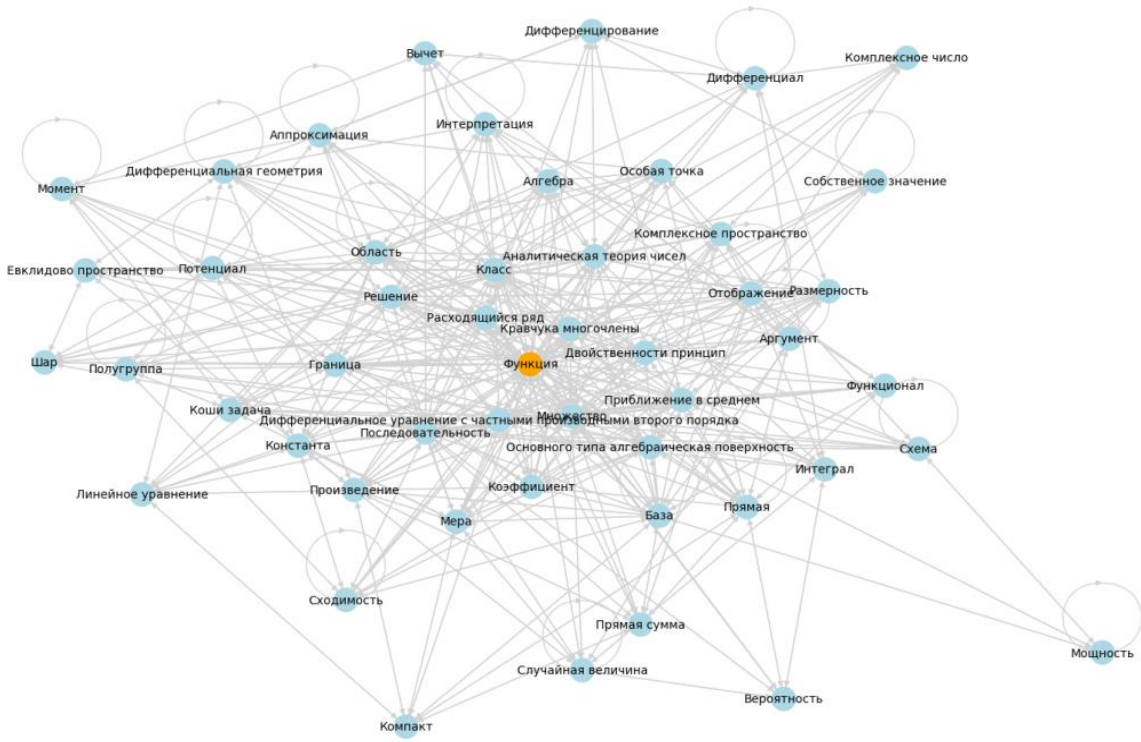

Рис. 2. Эко-граф узла «Функция» (подграф KG SciLibRu).

Качественный анализ привел к достраиванию графа и онтологии библиотеки.

1.3. Навигация по данным SciLibRu

Навигация по данным SciLibRu реализована с помощью KG. Вершины KG (статьи, термины, формулы SjD) – это экземпляры элементов онтологии, связи – это связи тезауруса SjD. Для описания онтологии использованы следующие формальные определения.

Объекты и ресурсы:

OB={O1,...,On}, where Oi, *i=1,..,n, – множество* информационных объектов;

R={r1,...,rm} , где *rj*, *j=1,…,m, – множество* информационных ресурсов (типы информационных объектов);

TYPE(O)=r – отношение, где каждому информационному объекту *Oi* соответствует информационный ресурс *rj*;

IsRe(r1,r2) – отношение иерархии ресурсов, где *r2* ($r2 \geq 0$, $r2 \subset r1$) составляет часть ресурса *r1* (подресурс).

Атрибуты:

A = {a1,…,ak} – множество атрибутов информационных ресурсова, *Z*(*ai*) – значение атрибута *ai*.

Источники модели данных:

SX = {(rx,Ar)}, – множество пар, где *rx* и *Ar* – ресурсы и их атрибуты,

G(rx,Ar) = (r,A), G – отношение и функция сопоставления элементов исходной модели данных с информационными ресурсами информационной системы и их набором атрибутов.

Данные источников: *X = (SX,G)*.

Предметная область тезауруса *Th*: *Th = (P,T,R)*, где *P – множество* концептов, *T – множество* вербальных терминов концептов, *R – множество* вертикальных и горизонтальных связей терминов и концептов.

F – функция сопоставления каждого информационного объекта из *OB* с соответствующими терминами из *T*: *F(oi) = {t}, t∈T, oi∈OB.*

Тезаурус *Th* определен для онтологии семантической библиотеки на языке OWL.

KG определяется на основе RDF-схемы как триплет (*s,p,o*).

Каждая тройка представляет собой упорядоченный набор терминов RDF:

субъект $s \in U \cup B$, предикат $p \in U$ и объект $o \in U \cup B \cup L$.

Кроме того, термин RDF – это URI, $u \in U$, «пустой» узел $b \in B$ или литерал $l \in L$. Значит, *KG* лучше всего представляется на основе схем RDF, но не каждое представление данных RDF рассматривается как *KG*.

Основные связи, которые реализованы между элементами *KG*:

- объект ↔ объект;
- концепт ↔ концепт;
- объект ↔ концепт ↔ объект;
- концепт ↔ объект ↔ концепт;
- классификатор ↔ концепт ↔ классификатор;
- концепт ↔ классификатор ↔ концепт;
- объект ↔ классификатор ↔ объект;
- классификатор ↔ объект ↔ классификатор.

1.4. Применение KG SciLibRu

Множество данных в виде KG имеет большое значение для создания рекомендательных систем с применением LLM для коммуникации. Одну из актуальных проблем LLM можно сформулировать как *ограничение знаний, на которых формируется ответ* LLM *при обеспечении полноты данных описания* SjD. Вариантом решения этой задачи может быть использование KG SjD в качестве входного множества данных для LLM, ранее предобученной на общих данных.

В нашем исследовании были проведены эксперименты [18] применения LLM к работе с SjD ODE в библиотеке SciLibRu. Разработана методология взаимодействия LLM и KG SjD ODE на основе инструкций, применяемых к описанию SjD в виде KG. На примере SjD ODE мы ограничиваем знания модели KG SjD, не позволяя ей выходить за границы SjD ODE. На выходе получаем ответы на естественном языке, опираясь на знания, представленные в библиотеке в виде KG.

В процессе экспериментов были выявлены следующие ошибки, которые влияют на качество данных в библиотеке SciLibRu.

1. Синтаксические ошибки в запросах. Это ошибки в запросах, генерируемых моделью LLM. Например, лишние «слэши» в предикатах, отсутствие префиксов. Для их исправления добавлена инструкция *промпт*.

2. Синтаксические ошибки в онтологии. Они вызывают у модели LLM сложности «в понимании», если классы и свойства некорректно именованы или аннотированы. Для их исправления в онтологию добавлены rdfs:label и rdfs:comment и учтены в дальнейшем при доработке модели данных.

С учетом этих ошибок были сформулированы инструкции к LLM с использованием модели KG SjD ODE для формирования SPARQL-запросов по запросам пользователя на естественном языке [18, 19].

2. ОБНОВЛЕНИЕ ДАННЫХ И НАПОЛНЕНИЕ БИБЛИОТЕКИ

Обновление данных и дальнейшее наполнение библиотеки, достраивание онтологии и графа знаний реализуются при *интеграции научных статей*, которые проходят предобработку для загрузки в библиотеку. В процессе предобработки данных был разработан *гибридный подход* для интеграции слабоструктурированных данных, содержащих символьную информацию (формулы в LaTeX). Были выявлены проблемы при реализации подхода, которые решались программным путем, созданием скриптов на Python.

Этап предобработки исходных данных журнальных статей для наполнения библиотеки SciLibRu потребовал анализа как самих данных, так и методов конвертации из формата LaTeX в стандартизированный формат XML, соответствующий заданным спецификациям.

2.1. Первичный анализ слабоструктурированных данных

Архивы выпусков выбранных периодических научных изданий хранятся в формате «ГГГГ-Н.rar» (где ГГГГ – год, Н – номер), где обязательно присутствует главный исходный файл – IPI.TEX. С помощью файла IPI.TEX выполняется формирование выпуска, при этом редакциями используются существующие шаблоны, обозначения и авторская стилистика применения LaTeX-макрокоманд.

Для выбора стратегии предобработки разнородных составных архивов были проанализированы некоторые подходы и методы, доступные при заданных (ограниченных по времени и мощности) вычислительных ресурсах.

В качестве стратегии для предобработки рассматривались:

А. Синтаксический анализ (парсинг, parsing) – построение формальной грамматики и использование парсер-генераторов для создания анализатора исходного текста. Этот подход требует тщательной проработки грамматики и сложен в реализации из-за гибкости и расширяемости LaTeX.

Б. Переопределение макрокоманд – модификация стандартных и пользовательских команд LaTeX для генерации XML-разметки вместо типографского вывода. Этот метод позволяет сохранить семантику документа, но требует глубокого понимания системы макросов LaTeX.

В. Использование промежуточных форматов – конвертация через промежуточные форматы (DVI, PDF) с последующим извлечением структуры и содержания. При этом подходе может теряться семантическая информация, но он проще в реализации.

Г. Гибридные решения – комбинация различных подходов, например, использование парсера для базовой структуры документа и переопределение макрокоманд для специфических конструкций.

Д. Использование нейросетевых моделей – для трансформации исходного текста LaTeX в XML. Этот подход позволяет достичь высокой точности конвертации, но требует значительных вычислительных ресурсов и большого объема данных для обучения модели.

Надо отметить, что самый качественный и универсальный подход состоит в использовании нейросетевых моделей, но высокие требования к вычислительным ресурсам и объему данных для обучения модели усложняют его применение в данном случае.

Использование промежуточных форматов позволяет частично упростить обработку, однако возникают проблемы при работе с устаревшими форматами документов. Кроме того, возникают накладные расходы на хранение и обработку промежуточных форматов.

Рассмотрены также методы и инструменты преобразования документов по извлечению структурированной информации [20, 21] из неструктурированных или полуструктурированных текстов, что имеет прямую аналогию с задачей преобразования LaTeX-разметки в семантически обогащенный XML.

E. Изучены универсальные конвертеры и программатические подходы (programmatic approach). Проведены исследования по созданию *кастомных парсеров* для LaTeX с использованием регулярных выражений, синтаксических анализаторов (например, ANTLR) или специализированных библиотек для работы с текстовыми данными. Эти исследования позволили оценить сложность разработки собственного решения и его преимущества в плане точности и контроля выходного формата.

В результате проведенного анализа методов (A-E) сделаны выводы, что наиболее простым и надежным подходом представляется *гибридный*: *комбинация парсинга для базовой структуры документа и переопределения макрокоманд для специфических конструкций и пользовательских команд*.

2.2. Разработка архитектуры преобразователя исходного документа и построение соответствующего XML-документа

Разработка системы конвертации LaTeX в XML реализована путем создания лексера (лексического анализатора) и парсера (синтаксического анализатора), авторского гибридного «компилятора» [22–24].

Сначала был проведен эксперимент: была произведена тестовая сборка файлов с помощью существующих систем сборки LaTeX. Далее, несмотря на наличие PDF-файлов, собранных в архиве, было принято решение провести тестовую сборку для проверки возможности использования решения, основанного на генерации промежуточного файла.

2.2.1. Проблемы кодировки

В ходе сборки были выявлены проблемы с кодировкой. Кодировка файлов произведена в CP866 (также известна как IBM866), разработана в паре с основной кодировкой (с которой совпадает по набору символов) в середине 1980-х годов в Вычислительном центре Академии наук СССР [25]. Эта кодировка пользовалась большой популярностью среди советских пользователей IBM PC-совместимых ПК.

Все кириллические символы архивов оказались нечитаемыми в исходном виде, в связи с чем появилась задача перекодировки файлов в кодировку UTF-8, как основную систему кодировки документов для подавляющего большинства повседневных задач для текстов в LaTeX. Для решения проблемы преобразования

файлов был написан скрипт на языке Python, изменяющий кодировку с сохранением исходного файла с расширением «.backup». В ходе работы выяснилось, что существующие механизмы обнаружения кодировок (разные версии uchardet) не определяют CP866 корректно (во всех случаях кодировка определена как турецкий подтип ISO). Это связано с тем, что CP866 – единственная или одна из немногих IBM-образных таблиц, получивших достаточно широкое распространение и при этом не вошедших в спецификацию ISO.

2.2.2. Проблемы специфических библиотек

Некоторые библиотеки (в частности, «acad.sty») не содержатся в основных репозиториях и пакетах поставки LaTeX-систем на Linux-дистрибутивы, а CTAN (архив Comprehensive TeX Archive Network) более не считает пакет активным. В результате эксперимента с тестовой сборкой файлов было принято решение разработки собственного лексера-парсера (компилятора) и последующего траверса AST (Abstract Syntax Tree). Были выбраны существующие генераторы грамматик (ANTL4 + Python). Разделение на лексер и парсер не осуществлялось, единая грамматика была сформирована с целью упрощения организации правил и управления конфигурацией.

2.3. Реализация прототипа конвертера, удовлетворяющего заявленным требованиям

Для работы был использован пример формата XML, конвертацию в который требовалось осуществить. С целью облегчения и ускорения отладки были составлены *модельные примеры корректных входного и выходного файлов*.

В процессе анализа исходных файлов для построения модельного файла был также решен ряд проблем с пользовательскими макрокомандами и грамматиками.

2.3.1. Исследование пользовательских макрокоманд

Все (или почти все) файлы с «полезным содержимым», то есть декларативным LaTeX-кодом, содержат определения пользовательских макрокоманд. Кроме того, нередко встречается переопределение макрокоманд, при этом не всегда с сохранением смысла. В самом файле IPI.TEX количество и содержание переопределений также нестабильны и изменяются от выпуска к выпуску. Веро-

ятнее всего, это связано с различной тематикой выпусков (например, определение *D* для быстрого начертания *дисперсии* для выпуска, посвященного теории вероятностей).

Названная проблема существенно осложняет разработку компиляторного решения и в очередной раз указывает на удобство нативного (оптимизированного для конкретного случая) решения (использования существующих систем сборки для генерации промежуточного или конечного файла и последующей обработки такового).

Для устранения несогласованности пользовательских макрокоманд был написан Python-скрипт, который в два прохода по файлам производит раскрытие пользовательских макросов (запись – раскрытие: 2 прохода). При этом пользовательские макросы (которые определены в файлах статей) имеют приоритет над глобальными. Такое действие относится к предобработке данных и выполняется до запуска непосредственного решения, аналогично с программой смены кодировки.

2.3.2. Результат исследований пользовательских макрокоманд

Написанный прототип успешно конвертировал модельный исходный файл в модельный целевой файл, опираясь на очевидные семантические маркеры (например, форматированная строка «Доказательство» для отделения доказательства присутствовала во всех файлах в том или ином виде.)

Грамматика модельного файла занимала 107 строк без учета сопроводительных пометок. В ходе реализации грамматики для «рабочего», а не модельного файла, объем грамматики быстро вырос до ~400 строк, а ANTLR-генерированные Python-программы стали нестабильными. Появились ошибки, например разбор текстового сегмента, как правило, после перевода на новую строку (символ абзаца в LaTeX – «\\», при этом LaTeX-правила начинаются с «\»).

2.3.3. ANTLR4 и сложные грамматики

В ходе работы над практической грамматикой установлено, что фреймворк ANTLR4 представляет собой мощный и расширяемый генератор парсеров по заданным правилам. Однако конфигурация ANTLR4 строго формализована и не допускает ручного управления и временного отхода от строгих правил реализации

грамматических правил обработки языка, что осложняет разработку практических лексеров и парсеров. В частности, при реализации вручную рекурсивного спуска можно реализовать «наивный lookahead» путем поиска закрывающего тэга с углублением рекурсии при обнаружении повторного открывающего. С ANTLR4, как и любым другим генератором, это невозможно.

Несмотря на отличную работу ANTLR4-генератора на модельных файлах, попытка применить его на практике столкнулась с проблемой *конфликтующих правил*, а также нетривиальной приоритетной очередности составленных выражений. Когда грамматика для практического файла достигла ~400 строк и ее отладка стала проблематичной, стало очевидно, что использование функционала грамматических генераторов необходимо пересмотреть.

Использование ANTL4 сопряжено с необходимостью считаться с особенностями подхода.

2.4. Оценка эффективности предложенного подхода путем экспериментального тестирования и сравнения результатов с существующими аналогичными инструментами

В исследовании был реализован траверс существующего дерева для формирования AST с применением лексера и парсера LaTeX для более корректного и качественного разбора текста. Были рассмотрены основные популярные приложения: pdftex, xetex, luatex, bibtex, miktex.

В ходе анализа выяснилось, что как такового построения AST они не производят. Основными недостатками оказались:

- неспособность использовать семантические маркеры (не-LaTeX команды) в качестве управляющих символов для правил без вмешательства в систему разбора исходных файлов;

- высокий порог вхождения для корректной настройки шаблонов и собственных макросов;

- исходный файл должен быть представлен в виде отдельного документа, не имеющего зависимостей.

После экспериментов построения конфигурации была реализована рабочая система раскрытия большинства макросов методом двойного прохода по файлу. Идея заключается в следующем: объединить все статьи в единый выходной

LaTeX-файл, в котором раскрыть все макрокоманды с учетом приоритета и области видимости. Это позволило избежать сложностей в виде разрешения зависимостей и раскрытия пользовательских макрокоманд.

Реализация алгоритма состояла из трех этапов: раскрытие макросов, разрешения подключений зависимостей и объединение текста.

ЗАКЛЮЧЕНИЕ

В результате исследования была проведена оценка качества данных библиотеки SciLibRu, выявлены направления для их совершенствования как на этапе загрузки (предобработки данных с помощью гибридного подхода), так и на этапе достраивания онтологии и KG. Было проведено обширное исследование методов и подходов предобработки сложно структурированных файлов и выработан подход для их конвертации и интеграции в семантическую библиотеку. Получен набор структурированных данных, который может быть использован для интеграции LLM с KG SciLibRu для коммуникации на естественном языке.

Дальнейшие исследования направлены на улучшение качества данных SciLibRu с целью формирования датасета для работы LLM с русскоязычными научными текстами.

Благодарности

Работа выполнена в рамках выполнения темы НИР «Математические методы анализа данных и прогнозирования» ФИЦ ИУ РАН.

СПИСОК ЛИТЕРАТУРЫ

1. *Серебряков В.А., Атаева О.М.* Информационная модель открытой персональной семантической библиотеки LibMeta // Научный сервис в сети Интернет: труды XVIII Всероссийской научной конференции (19–24 сентября 2016 г., г. Новороссийск). М.: ИПМ им. М.В. Келдыша, 2016. С. 304–313.
URL: http://keldysh.ru/abrau/2016/3.pdf.

2. *Rospocher M., Tonelli S., Serafini L.and Pianta E.* Corpus-based terminological evaluation of ontologies // Applied Ontology. 2012. Vol. 7, No. 4. P. 429–448. https://doi.org/10.3233/AO-2012-0114

3. *Ataeva O., Serebryakov V., Tuchkova N.* Ontological approach to a knowledge graph construction in a semantic library // Lobachevskii J. of Mathematics. 2023. Vol. 44, No. 6. P. 2229–2239. https://doi.org/10.1134/S1995080223060471

4. Handbook on Ontologies. Editors: Steffen Staab, Rudi Studer, Springer-Verlag Berlin Heidelberg, 2004. https://doi.org/10.1007/978-3-540-24750-0

5. *Атаева О.М., Серебряков В.А., Тучкова Н.П.* Подходы к организации математических знаний при формирования предметных тезаурусов различных разделов математики // CEUR Workshop Proceedings. 2018. Vol. 2260. P. 42–54. https://doi.org/10.20948/abrau-2018-66

6. *Hlomani H., Stacey D.,* Approaches, methods, metrics, measures, and subjectivity in ontology evaluation: A survey // Semantic Web Journal. 2014. Vol. 1, No. 5. P. 1–11. https://www.semantic-web-journal.net/system/files/swj657.pdf

7. *Lozano-Tello A. and Gómez-Pérez A.* Ontometric: A method to choose the appropriate ontology // Journal of Database Management. 2004. Vol. 15, No. 2. P. 1–18. https://doi.org/10.4018/jdm.2004040101

8. *Шрейдер Ю.А.* Тезаурусы в информатике и теоретической семантике // Научно-техническая информация. Сер. 2. 1971. № 3. С. 21–24.

9. *Лукашевич Н.В.* Тезаурусы в задачах информационного поиска. М.: Изд-во МГУ, 2011. 495 с.

10. *Харари Ф.* Теория графов. Пер. с англ. и предисл. В.П. Козырева. Под ред. Г.П. Гаврилова. Изд. 2-е. М.: Едиториал УРСС, 2003. 296 с.

11. *Barrasa J., Webber J.* Building Knowledge Graphs: A Practitioner's Guide. O'Reilly. 2023. 290 p.

12. *Biswas G., Bezdek J., Oakman R. L.* A knowledge-based approach to online document retrieval system design // In Proc. ACM SIGART Int. Symp. Methodol. Intell. Syst. (ISMIS '86) 1986. P. 112–120. https://doi/10.1145/12808.12821

13. *Гаврилова Т.А., Кудрявцев Д.В., Муромцев Д.И.* Инженерия знаний. Модели и методы: Учебник. СПб.: Изд-во «Лань», 2016. 324 с.

14. *Pan S. et al.* Unifying Large Language Models and Knowledge Graphs: A Roadmap // in *IEEE Transactions on Knowledge and Data Engineering*. 2024. Vol. 36, No. 7. P. 3580–3599. https://doi.org/10.1109/TKDE.2024.3352100

15. *Luo L. et al.* Graph-constrained reasoning: Faithful reasoning on knowledge graphs with large language models // arXiv preprint arXiv:2410.13080. 2024. https://doi.org/10.48550/arXiv.2410.13080

16. *Виноградов И.М.* (Гл. ред.). Математическая энциклопедия (в 5 томах)/ М.: Советская энциклопедия (1977–1985).

17. *Фаддеев Л.Д.* (Гл. ред.). Энциклопедия математической физики. Энциклопедия. М.: Большая русская энциклопедия, 1998. 692 с.

18. *Ataeva O.M., Tuchkova N.P.* Adaptation of the language model for mathematical texts in the semantic library // System Informatics (Системная информатика). 2025. No. 27. P. 59–75.

19. *Будзко В.И., Атаева О.М., Тучкова Н.П.* Автоматизация доступа к информации при навигации по данным семантической библиотеки и интеграции графа знаний с языковой моделью // Системы высокой доступности. 2025. Т. 21. № 2. С. 5−11. https://doi.org/10.18127/j20729472-202502-0

20. *Клюкин А.А., Широков А.А.* Автоматизированная система подготовки слабоструктурированной информации [Электронный ресурс] // Гаудеамус. 2014. №2 (24).
URL: https://cyberleninka.ru/article/n/avtomatizirovannaya-sistema-podgotovki-slabostrukturirovannoy-informatsii (дата обращения: 01.11.2025).

21. *Куртюкин С.В.* Метод автоматизированного формирования сборников архивных документов [Электронный ресурс] // Теория и практика современной науки. 2018. №5 (35).
URL: https://cyberleninka.ru/article/n/metod-avtomatizirovannogo-formirovaniya-sbornikov-arhivnyh-dokumentov (дата обращения: 01.11.2025).

22. *Ахо А., Сети Р., Ульман Дж.* Компиляторы: принципы, технологии, инструменты. М.: Вильямс, 2001, 762 с.

23. *Волкова И.А., Вылиток А.А., Руденко Т.В.* Формальные грамматики и языки. Элементы теории трансляции: учебное пособие для студентов II курса М.: Изд-во Моск. гос. ун-та, 2009.

24. *Гладкий А.В.* Формальные грамматики и языки. М.: Наука, Гл. ред. физ.-мат. лит., 1973, 368 с.

25. *Брябрин В.М., Ландау И.Я., Неменман М.Е.* О системе кодирования для персональных ЭВМ // Микропроцессорные средства и системы. 1986. № 4. С. 61–64.

Scilibru, The Library Of Scientific Subject Domains

O. M. Ataeva[1] [0000-0003-0367-5575], N. P. Tuchkova[2] [0000-0001-5357-9640],
K. B. Teymurazov[3] [0009-0000-0876-4222], A. Abdyshov[4] [0009-0000-4971-0378],
M. G. Kobuk[5] [0009-0002-9834-8218]

[1, 2, 3]Federal Research Center "Informatics and Control" of the Russian Academy of Sciences, Moscow, Russia
[1, 4, 5]S. Y. Witte Moscow University, Moscow, Russia

[1]OAtaeva@frccsc.ru, [2]NTuchkova@frccsc.ru, [3]KTeymurazov@frccsc.ru,
[4]abdysovajdin@gmail.com, [5]mgkobuk@gmail.com

Abstract

The work is devoted to the problem of data integration for representing scientific subject areas based on their semantic description in the SciLibRu digital library. The LibMeta library's ontology and knowledge graph are used as the data model. SciLibRu is populated by adding data from scientific journals. The paper demonstrates how the stages of processing semi-structured scientific publications for their integration into the library's ontology are implemented. Completing all data preprocessing stages yields a dataset that can be used to train language models for queries in Russian-language scientific subject areas.

Keywords*: applied ontology, knowledge graph, data sources, analysis of semi-structured scientific publications.*

REFERENCES

1. *Serebryakov V.A., Ataeva O.M.* Informacionnaya model' otkrytoj personal'noj semanticheskoj biblioteki LibMeta // Nauchnyj servis v seti Internet: trudy XVIII Vserossijskoj nauchnoj konferencii (19–24 sentyabrya 2016 g., g. Novorossijsk). M.: IPM im. M.V. Keldysha, 2016. S. 304–313. URL: http://keldysh.ru/abrau/2016/3.pdf (In Russ.)

2. *Rospocher M., Tonelli S., Serafini L., Pianta E.* Corpus-based terminological evaluation of ontologies // Applied Ontology. 2012. Vol. 7, No. 4. P. 429–448. https://doi.org/10.3233/AO-2012-0114

3. *Ataeva O., Serebryakov V., Tuchkova N.* Ontological approach to a knowledge graph construction in a semantic library // Lobachevskii J. of Mathematics. 2023. Vol. 44, No. 6. P. 2229–2239. https://doi.org/10.1134/S1995080223060471

4. Handbook on Ontologies. Editors: Steffen Staab, Rudi Studer, Springer-Verlag Berlin Heidelberg, 2004. https://doi.org/10.1007/978-3-540-24750-0

5. *Ataeva O., Serebryakov V., Tuchkova N.* Podhody k organizacii matematicheskih znanij pri formirovaniya predmetnyh tezaurusov razlichnyh razdelov matematiki // CEUR Workshop Proceedings. 2018. Vol. 2260. P. 42–54. https://doi.org/10.20948/abrau-2018-66 (In Russ.)

6. *Hlomani H., Stacey D.* Approaches, methods, metrics, measures, and subjectivity in ontology evaluation: A survey // Semantic Web Journal. 2014. Vol. 1, No. 5. P. 1–11. https://www.semantic-web-journal.net/system/files/swj657.pdf

7. *Lozano-Tello A., Gómez-Pérez A.* Ontometric: A method to choose the appropriate ontology // Journal of Database Management. 2004. Vol. 15, No. 2. P. 1–18. https://doi.org/10.4018/jdm.2004040101

8. *Shrejder Yu.A.* Tezaurusy v informatike i teoreticheskoj semantike // Nauchno-tekhnicheskaya informaciya. Ser. 2. 1971. № 3. S. 21–24 (In Russ.).

9. *Lukashevich N.V.* Tezaurusy v zadachah informacionnogo poiska. M.: Izd-vo MGU, 2011. 495 s. (In Russ.).

10. *Harari F.* Teoriya grafov. Per. s angl. i predisl. V.P. Kozyreva. Pod red. G.P. Gavrilova. Izd. 2-e. M.: Editorial URSS, 2003. 296 s

11. *Barrasa J., Webber J.* Building Knowledge Graphs: A Practitioner's Guide. O'Reilly. 2023. 290 p.

12. *Biswas G., Bezdek J., Oakman R.L.* A knowledge-based approach to online document retrieval system design // In Proc. ACM SIGART Int. Symp. Methodol. Intell. Syst. (ISMIS '86), 1986. P. 112–120. https://doi/10.1145/12808.12821

13. *Gavrilova T.A., Kudryavcev D.V., Muromcev D.I.* Inzheneriya znanij. Modeli i metody: Uchebnik. SPb.: Izdatel'stvo «Lan'», 2016. 324 s. (In Russ.).

14. *Pan S. et al.* Unifying Large Language Models and Knowledge Graphs: A Roadmap // in *IEEE Transactions on Knowledge and Data Engineering.* 2024. Vol. 36, No. 7. P. 3580–3599. https://doi.org/10.1109/TKDE.2024.3352100

15. *Luo L. et al.* Graph-constrained reasoning: Faithful reasoning on knowledge graphs with large language models // arXiv preprint arXiv:2410.13080. 2024.

https://doi.org/10.48550/arXiv.2410.13080

16. *Vinogradov I.M.* (Gl. red.). Matematicheskaya enciklopediya (v 5 tomah) M.: Sovetskaya enciklopediya (1977–1985) (In Russ.).

17. *Faddeev L.D.* (Gl. red.). Enciklopediya matematicheskoj fiziki. Enciklopediya. M.: Bol'shaya russkaya enciklopediya.1998. 692 s. (In Russ.).

18. *Ataeva O.M., Tuchkova N.P.* Adaptation of the language model for mathematical texts in the semantic library // System Informatics. 2025. No. 27. P. 59–75.

19. *Budzko V.I., Ataeva O.M., Tuchkova N.P.* Access automation to information for navigating through semantic library data and integrating the knowledge graph with the language model // Highly Available Systems. 2025. V. 21. No. 2. P. 5−11. https://doi.org/ 10.18127/j20729472-202502-0. (In Russ.).

20. *Klyukin A.A., Shirokov A.A.* Avtomatizirovannaya sistema podgotovki slabostrukturirovannoj informacii. [Elektronnyj resurs] // Gaudeamus. 2014. Vol. 24, No. 2. URL: https://cyberleninka.ru/article/n/avtomatizirovannaya-sistema-podgotovki-slabostrukturirovannoy-informatsii (date of access: 01.11.2025) (In Russ.).

21. *Kurtyukin S.V.* Metod avtomatizirovannogo formirovaniya sbornikov arhivnyh dokumentov [Elektronnyj resurs] // Teoriya i praktika sovremennoj nauki. 2018. №5 (35). URL: https://cyberleninka.ru/article/n/metod-avtomatizirovannogo-formirovaniya-sbornikov-arhivnyh-dokumentov (data obrashcheniya: 01.11.2025).

22. *Aho A., Seti R., Ul'man Dzh.* Kompilyatory: principy, tekhnologii, instrument. M.: Vil'yams, 2001, 762 s. (In Russ.).

23. *Volkova I.A., Vylitok A.A., Rudenko T.V*. Formal'nye grammatiki i yazyki. Elementy teorii translyacii : uchebnoe posobie dlya studentov II kursa M.: Izd-vo Mosk. gos. un-ta,2009. (In Russ.).

24. *Gladkij A.V.* Formal'nye grammatiki i yazyki. M.: Nauka, Gl. red. fiz.-mat. lit., 1973, 368 s. (In Russ.).

25. *Bryabrin V.M., Landau I.Ya., Nemenman M.E.* O sisteme kodirovaniya dlya personal'nyh EVM // Mikroprocessornye sredstva i sistemy. 1986. № 4. S. 61–64 (In Russ.).

СВЕДЕНИЯ ОБ АВТОРАХ

АТАЕВА Ольга Муратовна – старший научный сотрудник Федерального исследовательского центра «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН), кандидат техн. наук, специалист в области системного программирования и баз данных.

Olga Muratovna ATAEVA – senior researcher at FRC SCS RAS, PhD, expert in the field of system programming and databases.

email: OAtaeva@frccsc.ru

ORCID: 0000-0003-0367-5575

ТУЧКОВА Наталия Павловна – старший научный сотрудник ФИЦ ИУ РАН, кандидат физ.-мат. наук. Специалист в области алгоритмических языков и информационных технологий.

Natalia Pavlovna TUCHKOVA – senior researcher at FRC SCS RAS, PhD in physics with a math degree. The expert in the field of algorithmic languages and information technologies.

email: NTuchkova@frccsc.ru

ORCID: 0000-0001-5357-9640

ТЕЙМУРАЗОВ Кирилл Борисович – ведущий программист ФИЦ ИУ РАН, специалист в области информационных технологий и баз данных.

Kirill Borisovich TEYMURAZOV – lead programmer at FRC SCS RAS, expert in the field of computer sciences and databases.

email: KTeymurazov@frccsc.ru

ORCID: 0009-0000-0876-4222

АБДЫШОВ Айдин – студент бакалавриата Московского университет имени С.Ю. Витте, область интересов: анализ данных и визуализация.

Aidin ABDYSHOV – undergraduate student at the S. Y. Witte Moscow University, area of interest: data analysis and visualization.

email: abdysovajdin@gmail.com

ORCID: 0009-0000-4971-0378

КОБУК Михаил Геннадьевич – студент бакалавриата Московского Университет имени С.Ю. Витте, область интересов NLP, анализ данных, системное программирование.

Mikhail Gennadievich KOBUK – undergraduate student at the S. Y. Witte Moscow University, area of interest: NLP, data analysis, systems programming.

email: mikhail.kobuk@mail.ru

ORCID: 0009-0002-9834-8218

Материал поступил в редакцию 10 ноября 2025 года