

УДК 004.89

**РАЗРАБОТКА АЛГОРИТМОВ ПАРСИНГА И ГЕНЕРАЦИИ
БИБЛИОГРАФИЧЕСКИХ ОПИСАНИЙ НАУЧНЫХ ПУБЛИКАЦИЙ
В СООТВЕТСТВИИ С ТРЕБОВАНИЯМИ ГОСТ**

Усачев М.К., студент гр. ПИБ-222, IV курс
Торопов М.Е., студент гр. ПИБ-222, IV курс
Научный руководитель: Великий В.А., ассистент
Кузбасский государственный технический университет
имени Т.Ф. Горбачева, г. Кемерово

В условиях цифровизации научной и образовательной деятельности возрастает объем публикуемых научных работ. Это обуславливает необходимость использования эффективных инструментов для их учета, систематизации и корректного оформления. Одним из ключевых элементов научной публикации является библиографическое описание, оформляемое в соответствии с требованиями государственных стандартов (ГОСТ) [1, 2].

На практике преподаватели и научные сотрудники сталкиваются с трудностями при составлении корректных библиографических описаний. Это связано как со сложностью и многообразием нормативных требований, так и с отсутствием удобных и универсальных инструментов автоматизации. Существующие программные решения зачастую не поддерживают актуальные требования ГОСТ, а также ограничены только определенным узким набором типов публикаций. Кроме того, большинство из них не предоставляет возможности обратного преобразования (парсинга) – извлечения структурированных данных из уже сформированного библиографического описания.

Исходя из этого, возникает проблема отсутствия универсальных инструментов, позволяющих одновременно осуществить генерацию и парсинг библиографических описаний с учетом требований ГОСТ и разнообразия типов научных публикаций. В связи с этим разработка алгоритмов автоматизированной генерации и парсинга библиографических описаний является актуальной научно-практической задачей.

Целью данного исследования является разработка алгоритмов и программного средства, обеспечивающих автоматизированную генерацию и парсинг библиографических описаний научных публикаций в соответствии с требованиями ГОСТ.

Для достижения поставленной цели решаются следующие задачи:

- анализ существующих решений и требований ГОСТ к оформлению библиографических описаний;
- разработка алгоритмов гибкой генерации и парсинга библиографических описаний и проектирование структуры базы данных для хранения сведений о публикациях;
- проведение верификации алгоритмов на тестовом наборе публикаций.

Научная новизна работы заключается в разработке универсальных алгоритмов, обеспечивающих двунаправленное преобразование данных: из структурированной формы в библиографическое описание и из текстового описания в структурированные данные.

Современные инструменты для формирования библиографических описаний можно условно разделить на несколько категорий: менеджеры библиографических ссылок, онлайн-генераторы и специализированные научные платформы.

К числу наиболее распространенных менеджеров ссылок относятся системы, такие как Zotero и Mendeley. Они предоставляют широкие возможности по хранению, систематизации и автоматическому формированию библиографических данных, а также интеграцию с текстовыми редакторами. Однако данные инструменты ориентированы преимущественно на международные стандарты оформления (APA, MLA, Chicago) и не обеспечивают полноценной поддержки российских стандартов ГОСТ.

Онлайн-генераторы библиографических описаний позволяют пользователю вручную вводить данные публикации и получать готовую запись. Несмотря на простоту использования, такие решения, как правило, реализуют ограниченное количество шаблонов и не учитывают особенности различных типов научных публикаций. Кроме того, подобные сервисы не поддерживают обратное преобразование библиографического описания в структурированный вид.

Существуют также научные электронные библиотеки, например, eLIBRARY, которые предоставляют автоматическую генерацию библиографических описаний на основе собственных баз данных. Такие системы обеспечивают более корректное соответствие требованиям ГОСТ, однако их функционал ограничен и не предполагает гибкой работы с пользовательскими данными, а еще не включает механизмов парсинга.

Анализ существующих решений показал, что ни один из рассмотренных подходов не обеспечивает одновременно:

- полноценную поддержку требований ГОСТ;
- универсальность для различных типов публикаций;
- возможность двунаправленного преобразования данных.

Структура библиографического описания строго регламентирована и зависит от типа публикации. Например, описание статьи из журнала включает сведения о названии журнала, номере выпуска и страницах, в то время как для монографий указываются издательство и место издания. При этом используются стандартизированные разделительные знаки и сокращения, что обеспечивает единообразие представления данных.

Дополнительной сложностью является вариативность представления отдельных элементов, допускаемая стандартом. Это затрудняет автоматическую обработку библиографических записей, особенно при решении задачи их парсинга.

Исходя из этого, требования ГОСТ накладывают ограничения как на структуру формируемых описаний, так и на алгоритмы их обработки, что необходимо учитывать при разработке программных решений.

В рамках решения поставленных задач спроектирована структура базы данных, обеспечивающая хранение сведений о научных публикациях, их авторах и параметрах библиографических описаний. Структура базы данных включает сущности, отражающие публикации, типы публикаций, авторов, а также динамические поля, состав которых определяется в зависимости от типа публикации. Такой подход позволяет обеспечить расширяемость системы и адаптацию к различным требованиям оформления библиографических описаний. Общая схема базы данных представлена на рисунке 1.

Главной особенностью разработанного решения является использование гибких алгоритмов генерации и парсинга библиографических описаний с применением большой языковой модели (LLM). В отличие от шаблонных подходов, такой метод учитывает вариативность оформления и может обрабатывать записи, которые не полностью соответствуют фиксированным шаблонам.



Рисунок 1 – Схема базы данных

В ходе проектирования мы разработали два алгоритма: генерации библиографического описания на основе структурированных данных и парсинга библиографического описания с обратным преобразованием (рисунок 2).

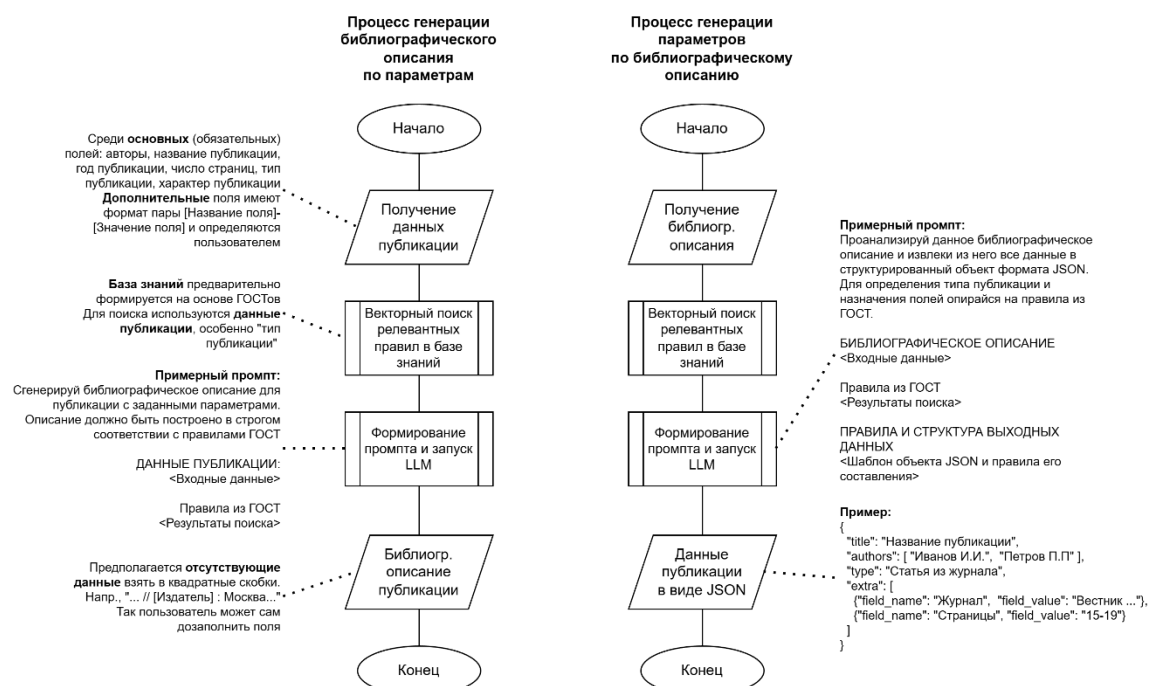


Рисунок 2 – Блок схемы алгоритмов генерации и парсинга

Алгоритм генерации библиографического описания на основе структурированных данных работает так. Сначала система получает данные о публикации из базы данных. Далее векторным поиском ищет правила оформления в базе знаний [3], содержащей требования ГОСТ и примеры библиографических записей. Это позволяет определить наиболее подходящий шаблон формирования описания с учетом типа публикации.

Дальше формируется текстовый запрос, который включает структурированные данные и найденные правила оформления. Затем он идет в большую языковую модель, которая генерирует строку библиографического описания.

Алгоритм парсинга библиографического описания реализует обратное преобразование и устроен похоже. На вход поступает текстовое библиографическое описание. Затем система векторным поиском ищет правила и шаблоны в базе знаний. После этого формируется запрос в LLM с исходным описанием и инструкцией по извлечению структурированных данных.

Модель возвращает JSON-объект, который содержит все основные элементы: авторов, название, источник публикации, год издания и другие поля. Полученные данные можно сохранить и использовать для дальнейшей обработки.

Предложенный подход позволяет отказаться от жестко заданных шаблонов и правил парсинга, что особенно важно когда записи оформлены по-разному. Использование векторного поиска обеспечивает выбор релевантных правил оформления, а применение языковой модели – гибкость и устойчивость к отклонениям в структуре входных данных.

Для верификации разработанных алгоритмов генерации и парсинга библиографического описания был взят набор из 10-ти публикаций, различаю-

щихся по типам. Для алгоритма генерации данные этих публикаций были вручную структурированы и переданы в виде JSON-объекта большой языковой модели GPT-5.3 вместе с документом ГОСТ и текстом задания. Анализ результатов выполнения LLM задачи выявил наличие ошибок в 3 из 10 библиографических описаний. Для алгоритма парсинга взяты уже готовые описания публикаций и аналогично загружены в LLM вместе с ГОСТ и заданием. В результате полученная JSON-структура полностью соответствует ожидаемой и синтаксически корректна в 10 из 10 случаев. Таким образом, отмечается высокая точность работы LLM даже без дообучения и комплексного промпт-инжиниринга.

В результате проведенного исследования разработаны алгоритмы генерации и парсинга библиографических описаний научных публикаций. Решение, основанное на использовании векторного поиска и больших языковых моделей, позволяет эффективно обрабатывать библиографические записи различной структуры и снижает зависимость от жестко заданных шаблонов. Это обеспечивает гибкость и расширяет область применения.

Список литературы:

1. ГОСТ 7.0.100-2018. Библиографическая запись. Библиографическое описание. Общие требования и правила составления. – М.: Стандартинформ, 2018. – 124 с.
2. ГОСТ 7.0.5-2008. Библиографическая ссылка. Общие требования и правила составления. – М.: Стандартинформ, 2008. – 19 с.
3. Бородулин, И. В. Увеличение точности больших языковых моделей с помощью расширенной поисковой генерации / И. В. Бородулин // Вестник науки. – 2024. – № 3 (72). – URL: <https://cyberleninka.ru/article/n/uvelichenie-tochnosti-bolshih-yazykovyh-modeley-s-pomoschyu-rasshirennoy-poiskovoy-generatsii> (дата обращения: 28.03.2026).