

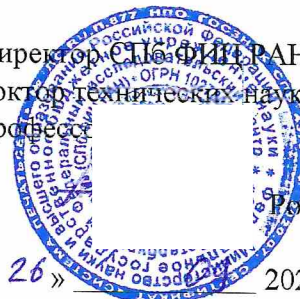
МИНОБРНАУКИ РОССИИ

**Федеральное государственное
бюджетное учреждение науки
«Санкт-Петербургский
Федеральный исследовательский центр
Российской академии наук»
(СПб ФИЦ РАН)**

14-я линия, д. 39, г. Санкт-Петербург, 199178
Тел.: (812) 328-33-11, факс: (812) 328-44-50,
e-mail: info@spcas.ru, web: http://www.spcas.ru
ОКПО 04683303, ОГРН 1027800514411,
ИНН/КПП 7801003920/780101001

УТВЕРЖДАЮ

Директор СПб ФИЦ РАН
доктор технических наук,
профессор



Ронжин А.Л.

« 26 » _____ 2025 года

«26» 09 2025 № 60/01-01-825

ОТЗЫВ ВЕДУЩЕЙ ОРГАНИЗАЦИИ

на диссертационную работу Шигарова Алексея Олеговича

**«Методы и инструментальные средства автоматизации процессов извлечения
данных из таблиц электронных документов неструктурированного формата»,
представленную на соискание ученой степени доктора технических наук по
специальности 2.3.5 – «Математическое и программное обеспечение
вычислительных систем, комплексов и компьютерных сетей»**

Актуальность темы диссертационной работы

Актуальность темы диссертационного исследования обусловлена: стремительным ростом объема табличных данных в электронных документах, их высокой востребованностью в различных прикладных областях и необходимостью автоматического структурирования для последующего использования. Документные таблицы представляют собой большой массив ценных фактов, которые могут быть использованы для ответа на запросы, наполнения баз знаний, генерации текстов на естественном языке и решения других задач. Однако, документные таблицы лишены явной семантики, что затрудняет их автоматическую обработку. Так называемое автоматизированное понимание таблиц заключается в восстановлении этой недостающей семантики для извлечения из них фактов в структурированном виде. Эта комплексная проблематика охватывает широкий спектр сложных задач — от извлечения таблиц из неструктурированных источников до их семантической интерпретации с привлечением внешних баз знаний.

Исследования в области автоматизированного понимания таблиц ведутся уже около трех десятилетий. Однако она до сих пор остается далеко не проработанной. Не обеспечена поддержка таблиц произвольной структуры, которые

характеризуются свободной компоновкой, сложной структурой ячеек и иерархичностью заголовков. Большинство существующих методов и программных средств автоматизированного понимания таблиц опирается на таксономии «стандартных» таблиц с атомарными ячейками и плоскими заголовками, что существенно сужает область их применимости. В связи с этим разработка методов и инструментальных средств автоматизации процессов извлечения данных из таблиц произвольной структуры представляет собой актуальную научную проблему. Её решение позволит более эффективно обрабатывать ключевые типы источников с включениями таблиц: печатно-ориентированные документы, рабочие книги (электронные таблицы) и др.

Научная новизна

Новизна проведенных исследований заключается в разработке новых моделей и методов автоматизированного понимания таблиц:

1. Унифицирована структура и согласована терминология задач автоматизированного понимания таблиц с учетом смежных научных направлений.
2. Предложен метод автоматизации процессов обнаружения и распознавания таблиц в нередактируемых печатных документах, базирующийся на использовании особенностей представления таблиц в языках описания страниц.
3. Создана модель представления табличных структур в процессах автоматизированного понимания таблиц.
4. Предложен метод автоматизации процессов анализа и интерпретации таблиц в рабочих книгах на основе выполнения пользовательских правил. Метод поддерживает таблицы произвольной структуры, включая свободную компоновку, объединенные ячейки и многоуровневые заголовки.
5. Создан проблемно-ориентированный язык правил для анализа и интерпретации таблиц, который позволяет пользователям разрабатывать собственные программы для извлечения данных из документов.
6. Разработан инструментарий для автоматизации ключевых процессов извлечения данных из таблиц произвольной структуры в электронных документах неструктурированных форматов.

Теоретическая и практическая значимость результатов исследования

Теоретическая значимость и научная новизна работы заключается в следующем:

1. Разработана детальная и комплексная структура задач автоматизированного понимания таблиц, которая превосходит существующие формулировки по своей полноте и является наиболее релевантной основой для проектирования прикладного программного обеспечения обработки табличных данных.
2. Предложен новый метод распознавания таблиц в нередактируемых документах, ключевым отличием которого является использование специфики

представления таблиц в языках описания страниц (PDL) для повышения точности сегментации и распознавания.

3. Создана модель представления таблиц, свободная от ограничений компоновочных типов, атомарности ячеек и плоских заголовков, что расширяет её применимость для анализа сложных и разнообразных табличных структур.

4. Предложен новый метод анализа и интерпретации таблиц в рабочих книгах, обеспечивающий поддержку произвольной структуры таблиц, включая свободную компоновку, структурированные ячейки и иерархические заголовки.

5. Создан специализированный язык программирования правил, который в отличие от языков общего назначения позволяет концентрироваться исключительно на логике извлечения данных, значительно упрощая процесс разработки.

6. Разработан комплекс инструментальных средств, который интегрирует в единую платформу: усовершенствованное распознавание таблиц из документов на основе PDL и гибкий анализ таблиц из рабочих книг на основе задаваемых пользователем правил.

Практическая значимость работы подтверждается внедрением результатов диссертации в ряде научных и промышленных проектов, где они были использованы для решения прикладных задач автоматизированного понимания таблиц:

1. В компании «CloudSDS Inc.» (США): для автоматического извлечения данных из паспортов безопасности химической продукции.

2. В компании «Mosaik Risk Solutions Inc.» (Индия): для анализа отчетов и оценки кредитных рисков.

3. В компании «ООО Адвентум Консалтинг»: для интеграции и консолидации табличных данных в сфере медиапланирования.

4. В ИДСТУ СО РАН: для автоматического создания электронных карт на основе отчетов Иркутскстата в интересах Правительства Иркутской области.

5. В ИДСТУ СО РАН: для построения онтологии экспертизы промышленной безопасности на основе табличных данных, предоставленных ООО «НТЦ «ИркутскийИхиммаш».

6. В ИДСТУ СО РАН совместно с Институтом национального развития Монголии: для наполнения информационно-аналитической системы статистическими данными о социально-экономическом положении районов Монголии.

7. В Гуанчжоуском университете (Китай): для разработки технологии кросс-контекстного обмена бизнес-документами.

8. В Пекинском технологическом университете (Китай): для создания эталонных наборов данных с целью тестирования производительности методов и обучения моделей распознавания таблиц.

Степень достоверности и апробация результатов.

Достоверность полученных результатов подтверждается представленными в диссертационной работе экспериментальными данными, качественным и количественным сравнением с имеющимися аналогами, а также программной реализацией. Разработанный комплекс инструментальных средств и материалы для воспроизведения проведенных экспериментов опубликованы в открытом доступе под свободными лицензиями. Достиженные результаты регулярно докладывались автором на всероссийских и международных конференциях, работа и результаты исследования внедряются в организации, связанные с тематикой данной научной работы, что подтверждено соответствующими актами.

Структура и содержание диссертационной работы

Диссертация состоит из введения, шести глав, заключения, списков сокращений, литературы, иллюстративного материала и таблиц, а также пяти приложений. Объем диссертации составляет 312 страниц, без приложений — 291; представлено 116 рисунков и 17 таблиц; процитировано 436 источников.

Во введении приводится общая характеристика диссертационной работы. Обосновывается актуальность выбранной темы, определяется научная новизна, теоретическая и практическая значимость основных результатов.

В первой главе представлен анализ проблематики и современного состояния исследований в области автоматизированного понимания таблиц. Вводятся ключевые понятия, связанные с документными таблицами, дается развернутая характеристика проблемного поля и предлагается структура для классификации соответствующих задач. Подробно рассматриваются постановки задач, существующие методы их решения, а также доступные инструменты для оценки — коллекции данных и тесты производительности. Определяются границы применимости существующих подходов, анализируются ограничения текущих исследований и формулируются перспективные направления для дальнейшей работы. Глава завершается краткими выводами.

Во второй главе рассматривается задача автоматизации распознавания таблиц в неотредактированных документах, ориентированных на печать. Вначале вводятся основные определения и предлагается модель представления страницы, которая составляет теоретическую основу исследования. Далее формулируется постановка задачи распознавания табличных структур. В качестве решения предлагается авторский метод.

Третья глава посвящена автоматизации анализа и интерпретации таблиц в рабочих книгах. В ней формулируются основные определения и предлагается модель документной таблицы, которая служит основой для дальнейшего исследования. Дается постановка задачи, для решения которой предлагается метод, основанный на пользовательском программировании правил. Подробно описываются состав

допустимых условий и действий в рамках этих правил, а также предметно-ориентированный язык для их описания (CRL). Определяется каноническая форма таблицы, предназначенная для представления результатов выполнения правил.

Четвертая глава представляет комплекс инструментальных средств, реализующих методы, изложенные в предыдущих главах. В ней описывается архитектура и функциональность разработанного программного обеспечения для распознавания, анализа и интерпретации таблиц. Особое внимание уделено автоматизации процесса трансферного обучения искусственных нейронных сетей для обнаружения таблиц. Подробно рассматривается применение CRL-правил на практических примерах, характерных для реальных задач. Приводятся результаты исследования пользовательского опыта работы с языком CRL, демонстрирующие его эффективность и удобство.

Пятая глава содержит результаты оценки производительности реализованных решений, а также их количественное и качественное сравнение с существующими аналогами. В начале главы представлены методики, использованные для проведения экспериментов и сравнений. Для оценки системы распознавания таблиц TabbyPDF приводятся результаты тестирования на общепризнанных бенчмарках: ICDAR-2013, SciTSR и IAIS. Для оценки системы анализа и интерпретации таблиц TabbyXL представлены результаты, полученные на специализированных тестовых наборах Troy200 и SAUS200, созданных на основе известных коллекций данных.

Шестая глава демонстрирует практическое применение методов, предложенных в диссертации, а также разработанного на их основе программного обеспечения для решения различных прикладных задач. В рамках главы рассматриваются следующие кейсы: анализ технической, финансовой и научной документации; интеграция данных государственной статистики; автоматизация процессов медиапланирования; построение онтологий предметных областей; кросс-контекстный обмен бизнес-документами.

В заключении перечисляются основные результаты диссертационной работы с кратким описанием их научной новизны. Приложения включают: грамматику языка CRL в расширенной форме Бэкуса–Наура; примеры таблиц из тестовых коллекций TROY200 и SAUS200; сопроводительные материалы прикладных задач четвертой главы; список подготовленных в рамках диссертационной работы Интернет-ресурсов; копии подтверждений внедрения результатов в индустриальной сфере.

Публикации по теме диссертации и личный вклад автора

Основные научные результаты диссертации опубликованы в 12 журналах, рекомендуемых Высшей аттестационной комиссией (ВАК) при Минобрнауки России для опубликования основных научных результатов диссертации на соискание ученой степени доктора наук по специальности 2.3.5 – «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей»

(технические науки). Все 12 изданий относятся к первой категории согласно итоговому распределению журналов Перечня ВАК по категориям в 2023 году, причем четыре из них — к первому уровню «Белого списка». Пять публикаций в журналах, рекомендуемых ВАК, подготовлены единолично. Также соискателем получено шесть свидетельств о государственной регистрации программ для ЭВМ.

Оценка содержания диссертации и её завершенности

Диссертация имеет ясную структуру, написана хорошим языком и представляет из себя законченное научное исследование. Содержание работы полностью соответствуют паспорту специальности 2.3.5 – «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей» (технические науки). Автореферат полно отражает содержание диссертации. Автореферат и диссертация оформлены в соответствии с актуальными требованиями. Подтверждена научная новизна и личный вклад соискателя в получение положений, выносимых на защиту.

Рекомендации по использованию результатов диссертации

Учитывая прикладной характер результатов диссертации, рекомендуется их дальнейшее внедрение в области обработки данных и знаний, представленных в виде произвольных таблиц электронных документов неструктурированного формата в следующих организациях, работающих на территории Российской Федерации:

1. В научных организациях по профилю исследований (ИДСТУ СО РАН, СПб ФИЦ РАН, Федеральном исследовательском центре информационных и вычислительных технологий и других) при исследовании и разработке программных систем обработки данных и знаний, представленных в виде произвольных таблиц электронных документов неструктурированного формата;

2. В органах государственного управления (в том числе: Правительстве Иркутской области, Росприроднадзоре) при формировании открытых государственных данных из документных источников табличной информации неструктурированного формата;

3. В индустриальных компаниях (например, ООО «Адвентум Консалтинг», ООО «НТЦ «ИркутскНИИхиммаш»), имеющих дело с массовой обработкой табличной информации, представленной в форматах Microsoft Excel, Google Sheets и пр.

Замечания по содержанию и оформлению работы

К диссертации и автореферату можно предъявить некоторые замечания:

1. Следовало шире посмотреть на проблему понимания документов неструктурированного формата, а также отразить возможность использования

полученных результатов для совершенствования больших языковых моделей GPT, DeepSeek и других.

2. Предложенные методы в наибольшей степени оптимизированы для документов с машиночитаемым текстом. В автореферате недостаточно освещены перспективы адаптации метода распознавания TabbyPDF для документов, требующих полноценного OCR (например, сканированных изображений), где PDL-специфичная информация отсутствует.

3. Автор указывает, что методы исследования основаны на применении эвристик, но при этом не конкретизирует применяемые эвристики и не сравнивает эффективность их применения между собой.

4. Несмотря на декларируемую простоту языка CRL, разработка эффективных наборов правил для таблиц со сложной, нерегулярной структурой по-прежнему требует значительной предварительно подготовки пользователя. В работе можно было бы более подробно рассмотреть вопросы разработки инструментов поддержки пользователя (например, визуальных конструкторов или средств автоматизированного тестирования правил) для дальнейшего снижения порога входа пользователя.

5. Имеются некоторые стилистические погрешности и опечатки в тексте работы.

Следует отметить, что приведенные замечания в целом не снижают научную и практическую ценность работы. Поставленные задачи в работе выполнены в полном объеме, цель исследования достигнута.

Заключение по диссертационной работе

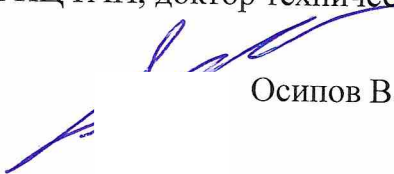
На основании анализа диссертации, автореферата и публикаций автора считаем, что диссертация А. О. Шигарова является завершенной научно-квалификационной работой, в которой решена актуальная научная проблема по разработке методов и инструментальных средств автоматизации процессов извлечения данных из таблиц с произвольной структурой, представленных в электронных документах неструктурированного формата, имеющая важное хозяйственное значение в области обработки данных и знаний.

Диссертационная работа соответствует требованиям п.п. 9–14 «Положения о порядке присуждения ученых степеней», утвержденным Постановлением Правительства Российской Федерации от 24 сентября 2013 г. № 842 (в имеющихся редакциях данного Постановления). Соискатель А. О. Шигаров заслуживает присуждения ученой степени доктора технических наук по специальности 2.3.5 – «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей».

Отзыв, диссертация и автореферат обсуждены на расширенном научном семинаре лабораторий автоматизации научных исследований, интегрированных

систем автоматизации и информационных технологий в системном анализе и моделировании СПб ФИЦ РАН 11 сентября 2025 года, протокол № 1.

Директор СПИИРАН СПб ФИЦ РАН, доктор технических наук, профессор



Осипов Василий Юрьевич

Главный научный сотрудник, руководитель лаборатории автоматизации научных исследований СПб ФИЦ РАН, доктор технических наук



Кулешов Сергей Викторович



Контактная информация организации:

Федеральное государственное бюджетное учреждение науки «Санкт-Петербургский
Федеральный исследовательский центр Российской академии наук»

Адрес: 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178, Россия

Сайт: <https://spcras.ru/>

Телефон: +7 (812) 508-33-11

E-mail: info@spcras.ru