

ОТЗЫВ
официального оппонента
на диссертацию 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

Таблицы являются естественным способом представления реляционных данных в электронных документах неструктурированных форматов (PDF, HTML, Excel и др.). Объем таких данных огромен. Они могут использоваться в критически важных приложениях поддержки принятия решений, интеллектуального анализа данных, построения баз знаний и пр. «Автоматизированное понимание таблиц» (АПТ) содержит такие основные этапы как распознавание, анализ и интерпретацию. В настоящее время эти процессы не могут выполняться полностью автоматически и требуют участия человека. Автоматизация обработки таблиц особенно важна в случаях массовой обработки технической, финансовой, научной документации, государственной статистики.

Вопросы АПТ исследуются уже более трех десятилетий, но общего решения этой проблемы до сих пор нет. Активно развиваются методы на основе глубокого обучения, векторных представлений, открытых данных. Общим недостатком современных методов АПТ является ограниченная поддержка небольшого числа наиболее распространенных таблиц компоновочных типов. При этом игнорируется большое количество различных специфических случаев компоновки таблиц, структурированность ячеек и иерархичность заголовков, что значительно сужает область применения существующих инструментов. Заявленная соискателем тема диссертационной работы актуальна в связи с необходимостью разработки методов и инструментов для автоматизированного извлечения данных из документных таблиц более произвольной структуры, что позволит преодолеть ограничения некоторых существующих решений и расширить возможности массовой обработки табличной информации в научных и прикладных областях.

Структура и содержание диссертации

Диссертация содержит введение, шесть глав, заключение, списки литературы,

сокращений и иллюстраций, а также пять приложений. Введение определяет цель и задачи, обосновывает актуальность темы, научную новизну и значимость работы. В Главе 1 приводится анализ совокупности задач АПТ, критический обзор методов АПТ с выявлением их ограничений и направлений дальнейшего развития, а именно использование PDL-специфики и поддержку произвольной структуры. В Главе 2 изложен метод автоматизации распознавания таблиц в неотредактируемых печатно-ориентированных документах (PDF), а также методика использования PDL-информации для сегментации страниц и извлечения физической структуры таблиц. В Главе 3 введен метод автоматизации анализа таблиц редактируемого формата (Excel), а также проблемно-ориентированный язык правил анализа и интерпретации таблиц CRL. В Главе 4 описаны инструментальные средства: TabbyPDF (для распознавания в PDF) и TabbyXL (платформа для анализа в Excel). В Главе 5 приведены экспериментальные данные, подтверждающие конкурентоспособность методов. В Главе 6 рассмотрено успешное применение методов в восьми научных и промышленных проектах (интеллектуальный анализ документов, интеграция данных и др.). В Заключении сформулированы выводы, подтверждающие достижение цели работы. Приложения содержат грамматику CRL, примеры таблиц и подтверждения внедрения.

Научная новизна, теоретическая и практическая ценность

Разработана совокупность новых научно обоснованных теоретических положений в области АПТ, включая следующие: усовершенствованная структура задач АПТ; метод автоматизации распознавания таблиц, научная новизна которого связана с применением PDL-специфики для повышения качества результатов; модель представления произвольных табличных структур; метод автоматизации анализа и интерпретации таблиц, научная новизна которого связана с поддержкой таблиц с произвольной структурой, описываемой пользовательскими правилами, программируемых на предложенном соискателем проблемно-ориентированном языке CRL; комплекс инструментальных средств, реализующий предложенные методы.

Теоретическая ценность работы состоит в решении важной научной проблемы автоматизации извлечения данных из документных таблиц с произвольной структурой, что имеет существенное социально-экономическое и хозяйственное значение в условиях цифровой трансформации экономики. Практическая ценность работы подтверждается созданием промышленно применимых инструментальных средств TabbyPDF и TabbyXL, внедрением результатов в реальные проекты обработки документов, использованием

в системах интеграции данных и управления знаниями, а также публикацией исходного кода под свободными лицензиями.

Соответствие диссертации критериям Положения о присуждении ученых степеней

В соответствии с **пунктом 9** Положения о присуждении ученых степеней, утвержденного постановлением Правительства Российской Федерации от 24.09.2013 № 842 (далее – Положение): «Диссертация на соискание ученой степени доктора наук должна быть научно-квалификационной работой, в которой на основании выполненных автором исследований разработаны теоретические положения, совокупность которых можно квалифицировать как научное достижение, либо решена научная проблема, имеющая важное политическое, социально-экономическое, культурное или хозяйственное значение, либо изложены новые научно обоснованные технические, технологические или иные решения, внедрение которых вносит значительный вклад в развитие страны».

Диссертация Шигарова А.О. полностью соответствует данному критерию, поскольку содержит следующие новые научно обоснованные технические и технологические решения:

- метод автоматизации распознавания таблиц в нередактируемых печатно-ориентированных документах на основе обработки информации специфичной для языков описания страниц;
- модель представления табличной структуры, поддерживающая произвольность компоновки, структурированность ячеек и иерархичность заголовков;
- метод автоматизации анализа и интерпретации таблиц на основе пользовательского программирования правил;
- проблемно-ориентированный язык правил анализа и интерпретации таблиц;
- комплекс инструментальных средств, реализующий предложенные методы.

Значительный вклад внедрения в развитие страны обоснован тем, что разработанный автором комплекс инструментальных средств (TabbyPDF, TabbyXL) применен в реальных проектах обработки документов, что подтверждается письмами о внедрении от промышленных компаний. Реализованные решения были применены, в том числе, в задачах обработки данных по государственной статистике (Иркутскстата), социально-экономического мониторинга (Института национального развития

Монголии), экспертизы промышленной безопасности («ИркутскНИИХиммаш»).

Предложенные в рамках диссертации технологии используются в научных проектах, поддерживаемых грантами РНФ, РФФИ, Совета по грантам Президента Российской Федерации и Министерства науки и высшего образования Российской Федерации. Инструментальные средства опубликованы под свободными лицензиями и используются научным сообществом (включая Институт национального развития Монголии, Пекинский технологический институт, Гуанчжоуский университет), а также зарубежными («CloudSDS», «Mosaik Risk Solutions») и российскими компаниями («Адвентум Консалтинг»).

В соответствии с приоритетами развития страны разработанные автором решения способствуют цифровой трансформации экономики через автоматизацию обработки документной информации, развитию технологий работы с большими данными и искусственным интеллектом, а также повышению эффективности систем государственного управления и бизнес-аналитики.

Таким образом, диссертация Шигарова А.О. представляет собой научно-квалификационную работу, в которой изложены новые научно обоснованные технические и технологические решения, внедрение которых вносит значительный вклад в развитие страны, что полностью соответствует требованиям пункта 9 Положения.

В соответствии с **пунктом 10** Положения отмечу следующее.

- Диссертация написана автором самостоятельно, обладает внутренним единством и содержит новые научные результаты, выдвигаемые для публичной защиты.
- Личный вклад автора подтверждается формулировками в автореферате и диссертации и включает постановку задач, разработку методов, моделей, языковых и инструментальных средств.
- Приведены сведения о практическом использовании результатов в пяти научных и трех индустриальных проектах, включая разработки компаний «CloudSDS», «Mosaik Risk Solutions» и «Адвентум Консалтинг».
- Предложенные технологические решения аргументированы и оценены в сравнении с известными аналогами в ходе экспериментальных исследований на тестовых коллекциях ICDAR-2013, SciTSR, IAIS, Troy200 и SAUS200. F-мера составила 92,1% на этапе распознавания и до 98,5% на этапе анализа и интерпретации. Качественно они превосходят аналоги за счет использования PDL-специфики и поддержки произвольной структуры таблиц.

В соответствии с **пунктами 11–13** Положения основные научные результаты диссертации опубликованы в 70 работах, включая:

- 12 статей в журналах, рекомендованных ВАК, относящихся к категории К1, а четыре статьи к первому уровню «Белого списка» РЦНИ и первому квартилю индекса SJR;
- 6 свидетельств о государственной регистрации программ для ЭВМ;
- 34 статьи в изданиях, индексируемых в международных базах данных научных изданий (Web of Science, Scopus).

Количество и качество публикаций полностью соответствуют требованиям для докторской диссертации по техническим наукам.

В соответствии с **пунктом 14** Положения в диссертации обеспечивается корректное цитирование и ссылки на источники заимствования материалов. Личный вклад автора в совместные работы четко обозначен.

Замечания

По представленной диссертационной работе имеется следующее замечание. Текст диссертации и автореферата содержит жаргонные термины, часто полученные побуквенным переводом английских терминов и, вообще говоря, понятные специалистам. Например, термин «растеризовать» на стр. 14 автореферата означает перевод информации в растровый вид. Некоторые термины носят неуклюжую форму. Среди них «каноникализация», означающая приведение данных к канонической форме. А «пробельное пространство» означает пространство пробелов.

Введенные таким способом термины, в том числе неуклюжие, с теоретической точки зрения в тексте определены однозначно, так что в ходе внимательного чтения диссертации не вызывают двоякого или неопределенного восприятия.

Отмеченное замечание не влияет на общую положительную оценку работы и не снижает научной и практической ценности полученных результатов.

Заключение

Диссертация Шигарова А.О. представляет собой завершенное научное исследование, соответствующее всем требованиям (пп. 9–14) Положения о присуждении ученых степеней, утвержденного постановлением Правительства Российской Федерации от 24.09.2013 № 842 «О порядке присуждения ученых степеней» (в редакции от 16.10.2024). Работа обладает научной новизной, теоретической и практической

директор Института вычислительного моделирования Сибирского отделения Российской академии наук – обособленного подразделения Федерального государственного бюджетного научного учреждения «Федеральный исследовательский центр «Красноярский научный центр Сибирского отделения Российской академии наук» (ИВМ СО РАН).

34

01 декабря 2025 г.

Институт вычислительного моделирования СО РАН – обособленное подразделение
Федерального государственного бюджетного научного учреждения «Федеральный
исследовательский центр «Красноярский научный центр Сибирского отделения
Российской академии наук», 660036, г. Красноярск, Академгородок, дом 50, стр. 44.
Телефон: +7 (391) 290-73-56, эл. почта: sek@icm.krasn.ru

Заместитель директора ИВМ СО РАН
по научной работе канд. техн. наук

Heard

С.В. Исаев

01 декабря 2025 г.

