Подход к проектированию системы подготовки научных статей в техническом университете

И. И. Максимов*

Предложен подход к подготовке научных статей с независимыми моделями содержания и требований издания. Конвейер объединяет разбор рукописи, формирование документов и контроль преобразования. Реализованный модуль создаёт проект LaTeX, а для входного DOCX — также редактируемый документ Word. Все 63 функциональных теста завершились успешно. Проверены обработка структурных элементов, сохранность объектов и обнаружение заданных нарушений. Подход позволяет повторно использовать обработчики при смене издания и выделять участки для редакторской проверки.

Ключевые слова: научная статья, обработка документов, структурирование, шаблон, LaTeX, контроль преобразования.

1. Постановка задачи

Рассматривается подготовка авторской рукописи по требованиям журнала или конференции. Для технического университета существенна возможность использовать одну систему при работе с разными изданиями. Переоформление должно сохранять текст, числовые данные, таблицы, рисунки, формулы и ссылки.

При анализе научных документов учитываются логическая структура [1], разметка объектов страницы [2], порядок чтения и иерархия элементов [3]. В настоящей работе структурный анализ включён в процесс переоформления рукописи и контроля сохранности материала.

Цель работы — разработать подход к построению системы подготовки статей с независимыми описаниями содержания и требований издания. На вход поступают рукопись и образец либо текстовые требования. Результат включает исходники для вёрстки и отчёт о проверке; для рукописи DOCX дополнительно формируется редактируемый документ Word.

* *Максимов Игорь Игоревич* — студент 3-го курса кафедры «Системы автоматизированной поддержки принятия решений» ФГБОУ ВО «Тамбовский государственный технический университет», e-mail: maksimov.i289@yandex.ru.

Maksimov Igor Igorevich — third-year student, Department of Automated Decision Support Systems, Federal State Budgetary Educational Institution of Higher Education “Tambov State Technical University”.

Опубликовано на условиях лицензии CC-BY 4.0
Published under CC-BY 4.0

2. Организация преобразования

Структурная модель рукописи хранит метаданные, разделы и абзацы. Таблицы, рисунки, формулы и списки представлены отдельными объектами с подписями, идентификаторами и ссылками. Сохраняются сведения об их положении в исходном файле. Профиль издания задаёт параметры страницы, полей, колонок, шрифта, заголовков и подписей, настройки класса LaTeX. Смена профиля изменяет оформление выделенных элементов без изменения алгоритма разбора рукописи.

Подход реализован в самостоятельном модуле проекта WEBTSTU. Анализатор входного пакета определяет основной файл и ресурсы. Обработчики DOCX, LaTeX и PDF формируют общую структурную модель. Для назначения роли блока учитываются три группы признаков: явная разметка (стили и нумерация DOCX, команды и окружения LaTeX), положение относительно соседних блоков и текстовые маркеры (например, номер раздела или подпись «Рисунок» либо «Таблица»). Явная структурная разметка задаёт первичного кандидата; при её отсутствии роль назначается при согласовании позиционных и текстовых признаков. Для PDF, где такая разметка обычно недоступна, используются две последние группы. Если признаки указывают на разные роли либо не дают однозначного решения, блок считается спорным: его содержимое сохраняется, а классификация требует редакторской проверки. Подключаемая языковая модель может предложить роль спорного блока, но не переписывает научный текст и не подменяет редакторское решение.

Генератор применяет профиль к структурной модели и создаёт проект LaTeX. При наличии компилятора формируется PDF. Редактируемый документ Word создаётся отдельно, только для входной рукописи DOCX, с использованием сохранённых объектов исходного файла. Для входных LaTeX и PDF получение документа Word не заявляется; обратное преобразование PDF в Word не выполняется. Промежуточные модели и результаты этапов сохраняются для проверки.

3. Контроль и результаты проверки

Структурная проверка сопоставляет количества таблиц, рисунков и формул, контролирует идентификаторы и связи. Содержательная проверка выявляет потери текста и объектов, включая строки и ячейки таблиц. Техническая проверка охватывает файлы, ресурсы, исходники LaTeX, сообщения компилятора, стили Word и ширину таблиц. Ошибки и предупреждения регистрируются раздельно; спорные случаи передаются на

редакторский контроль. Математическая эквивалентность формул этим способом не устанавливается.

Проверена версия 0.4.0 из архива WEBTSTU-main.zip в Python 3.13.5 и pytest 9.0.2. Внешняя языковая модель отключена. Запущены 63 теста из семи файлов: 22 — на разбор документов и назначение блоков, 32 — на формирование документов и сохранность объектов, 9 — на обработку пакетов и работу конвейера. Использованы искусственные документы с заданными свойствами. Успех означает выполнение ожидаемого условия, включая обнаружение специально внесённой ошибки.

Все 63 теста завершились успешно. Подтверждены сохранение номера и текста заголовка, отсутствие дублирования текста объединённых ячеек, перенос изображений, обработка двуязычного начального блока и обнаружение потери строк таблицы. Проверены создание исходников LaTeX и архива проекта, применение оформления и диагностика заданных нарушений.

Результаты относятся к проверенным программным компонентам. Число тестов не является числом обработанных статей. Этот запуск не оценивает точность распознавания на независимом корпусе, качество окончательной вёрстки PDF и работу внешней языковой модели.

4. Заключение

Реализован подход с независимыми моделями содержания и требований издания. Разбор, оформление и контроль объединены в единый конвейер. Функциональные тесты подтвердили выполнение заданных сценариев. Подход предназначен для повторного использования обработчиков при смене издания и контроля сохранности рукописи.

При подготовке текста и организации запуска тестов использовались инструменты на основе больших языковых моделей.

Список литературы

[1] A. Ramesh Kashyap, Y. Yang, M.-Y. Kan, “Scientific document processing: challenges for modern learning methods”, *International Journal on Digital Libraries*, **24**:4 (2023), 283–309. DOI: 10.1007/s00799-023-00352-7.

[2] A. Gemelli, S. Marinai, L. Pisaneschi, F. Santoni, “Datasets and annotations for layout analysis of scientific articles”, *International Journal on Document Analysis and Recognition*, **27**:4 (2024), 683–705. DOI: 10.1007/s10032-024-00461-2.

[3] J. Wang, K. Hu, Z. Zhong, L. Sun, Q. Huo, “Detect-order-construct: A tree construction based approach for hierarchical document structure analysis”,

Pattern Recognition, **156** (2024), 110836. DOI: 10.1016/j.patcog.2024.110836.

Статья поступила 7 октября 2026 г.

An approach to designing a scientific article preparation system for a technical university

I. I. Maksimov

An approach to scientific article preparation using independent models of content and publication requirements is proposed. The pipeline combines manuscript parsing, document generation, and validation. The implemented module produces a LaTeX project and, for DOCX input, an editable Word document. All 63 functional tests passed, covering structural processing, object preservation, and detection of predefined violations. The approach enables parsers to be reused across publications and identifies elements requiring editorial review.

Keywords: scientific article, document processing, structure, template, LaTeX, conversion validation.

References

[1] A. Ramesh Kashyap, Y. Yang, M.-Y. Kan, "Scientific document processing: challenges for modern learning methods", *International Journal on Digital Libraries*, **24**:4 (2023), 283–309. DOI: 10.1007/s00799-023-00352-7.

[2] A. Gemelli, S. Marinai, L. Pisaneschi, F. Santoni, "Datasets and annotations for layout analysis of scientific articles", *International Journal on Document Analysis and Recognition*, **27**:4 (2024), 683–705. DOI: 10.1007/s10032-024-00461-2.

[3] J. Wang, K. Hu, Z. Zhong, L. Sun, Q. Huo, "Detect-order-construct: A tree construction based approach for hierarchical document structure analysis", *Pattern Recognition*, **156** (2024), 110836. DOI: 10.1016/j.patcog.2024.110836.

Received on October 7, 2026