ООО «Амфител Плюс»
ОГРН: 1071690018036, ИНН: 1659071213, КПП: 165501001

Программное обеспечение «Textora»

Описание функциональных характеристик программного обеспечения

(для целей проведения экспертной проверки в Экспертном совете при Минцифры России)

г. Казань, 2026 г.

1. ОБЩИЕ СВЕДЕНИЯ

«Textora» - программное обеспечение для автоматизированного распознавания, сегментации и классификации документов физических лиц, а также для извлечения из них структурированных данных в корпоративных информационных системах организаций.

Система принимает на вход изображения или PDF-файлы документов, автоматически определяет тип документа, извлекает структурированные данные и возвращает результат через REST API в формате JSON. Обработка выполняется конвейером нейросетевых моделей в автоматическом режиме без участия оператора.

Поддерживаемые типы документов: паспорт гражданина Российской Федерации (страницы 2, 3, страницы регистрации 5-12), страховое свидетельство обязательного пенсионного страхования (СНИЛС).

Программное обеспечение разворачивается внутри контура заказчика и функционирует в изолированной локальной сети без доступа к сети Интернет. Интеграция с корпоративными информационными системами заказчика (ERP, СЭД, CRM и др.) осуществляется через REST API.

Типовые сценарии применения:

– автоматизация процессов KYC (Know Your Customer) в финансовых и страховых организациях;
– ввод данных из документов клиентов в системы документооборота и CRM;
– верификация личности при дистанционном обслуживании;
– автоматизация процессов авансовой отчётности и кадрового учёта.

2. ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ

Раздел содержит определения основных терминов, используемых в настоящем документе.

Термин	Определение
REST API	Программный интерфейс взаимодействия, основанный на архитектурном стиле REST (Representational State Transfer). Обеспечивает обмен данными между системами по протоколу HTTP.
JSON	Текстовый формат обмена данными (JavaScript Object Notation). Используется для передачи структурированных результатов обработки через REST API.
OCR	Оптическое распознавание символов (Optical Character Recognition) - технология перевода изображений рукописного или печатного текста в машиночитаемый формат.
MRZ	Машиночитаемая зона документа (Machine Readable Zone) - область паспорта, содержащая данные в стандартизированном формате для автоматического считывания.

Термин	Определение
PDF	Формат электронного документа (Portable Document Format), поддерживающий текстовое и графическое содержимое.
Docker	Платформа контейнеризации, обеспечивающая упаковку приложения и его зависимостей в изолированный контейнер для развёртывания в инфраструктуре заказчика.
FastAPI	Веб-фреймворк для разработки REST API на языке Python. Используется для реализации интерфейсного слоя системы.
YOLO	Семейство нейросетевых архитектур для детекции и сегментации объектов на изображениях (You Only Look Once).
ViTOCR	Нейросетевая архитектура на основе Vision Transformer, используемая в системе для оптического распознавания символов.
EdgeNeXt + SDTA	Нейросетевая архитектура, используемая в системе для классификации типов документов и определения их ориентации.
TorchScript	Механизм компиляции нейросетевых моделей PyTorch в статический граф для ускорения инференса.
SQLite	Встраиваемая реляционная база данных, используемая для хранения результатов обработки задач.
KYC	Know Your Customer - процедура идентификации и верификации личности клиента, применяемая в финансовых и страховых организациях.
ZIP-архив	Формат архивного файла, поддерживающий сжатие данных. Используется в системе для пакетной передачи нескольких изображений на обработку.
OpenAPI	Стандарт описания REST API. Документация в формате OpenAPI генерируется системой автоматически и доступна через Swagger UI и ReDoc.

3. ФУНКЦИОНАЛЬНЫЕ ВОЗМОЖНОСТИ

Раздел содержит описание функциональных возможностей программного обеспечения «Textora».

Приём и предобработка входных данных

Система обеспечивает приём документов в следующих форматах:

– изображения: JPEG, PNG, TIFF, BMP, WEBP;
– PDF-файлы (текстовые и сканированные), многостраничные;
– ZIP-архивы для пакетной обработки нескольких изображений.

В процессе предобработки выполняется:

– извлечение изображений из многостраничных PDF-файлов;

– автоматическая коррекция геометрических искажений: угла наклона, перспективных искажений;
– улучшение качества изображения для повышения точности распознавания.

Ограничения: максимальный размер входного файла - 50 МБ; максимальный размер очереди задач - 1000 единиц.

Классификация документов

Система обеспечивает:

– автоматическое определение типа документа с использованием нейросетевого классификатора;
– определение ориентации документа (0°, 90°, 180°, 270°) и автоматическую ротацию до стандартного положения;
– сегментацию области документа на изображении.

Распознавание и извлечение данных

Система обеспечивает распознавание печатного и рукописного текста на русском языке (кириллица), а также частично на английском языке (для обработки машиночитаемых зон).

Паспорт гражданина Российской Федерации

Из страницы 2 извлекаются:

– серия и номер;
– дата выдачи;
– орган выдачи («Кем выдан»);
– код подразделения;
– источник изображения (фото, скан, фото скана, фото экрана, селфи).

Из страницы 3 извлекаются:

– фамилия, имя, отчество;
– дата рождения;
– место рождения;
– пол;
– серия и номер;
– строки машиночитаемой зоны MRZ1 и MRZ2.

Из страниц 5-12 извлекаются блоки регистрации. Каждый блок содержит:

– номер страницы, индекс блока на странице;
– тип блока (печатный / рукописный);
– для печатных блоков: регион, район, населённый пункт, улица, дом, корпус, квартира, дата регистрации, подразделение, код подразделения, статус регистрации.

Страховое свидетельство обязательного пенсионного страхования (СНИЛС)

Извлекаются:

– страховой номер индивидуального лицевого счёта (11 цифр);
– фамилия, имя, отчество;
– дата рождения;
– пол;
– место рождения;
– дата регистрации в ПФР.

Валидация машиночитаемой зоны (MRZ)

Система выполняет автоматическую валидацию машиночитаемой зоны паспорта:

– проверку контрольных сумм;
– коррекцию типичных OCR-ошибок в полях MRZ;
– нормализацию кодов страны, пола и символов-заполнителей.

Постобработка результатов

Система обеспечивает:

– нормализацию форматов дат к единому стандарту (ДД.ММ.ГГГГ);
– очистку и нормализацию извлечённых текстовых полей;
– валидацию корректности извлечённых данных;
– формирование структурированного результата в формате JSON.

Проверка качества изображения

Система предоставляет асинхронную проверку качества входного изображения или PDF-файла независимо от задачи распознавания. Результат проверки включает:

– детекцию и сегментацию документа на изображении, классификацию его типа;
– детекцию информационных полей документа с признаком их наличия на изображении;
– классификацию источника изображения: фото, скан, фото скана, фото экрана, селфи;
– признак grayscale-изображения;
– детекцию дефектов изображения: пальцы, перекрывающие поля документа; засветы; выход документа за границы кадра; размытие; низкое разрешение;
– перечень полей документа, перекрытых пальцами или засветами.

Верификация паспортных данных

Система предоставляет синхронную верификацию паспортных данных без обработки изображения. Принимает JSON-объект с полями паспорта, возвращает детальный отчёт по каждой проверке (всего 33 проверки).

Группы проверок:

– обязательные поля - наличие всех требуемых полей;
– ФИО - корректность символов, длина, отсутствие недопустимых слов и аббревиатур;
– дата рождения - корректность числа, месяца, года, отсутствие будущей даты;
– дата выдачи - корректность числа, месяца, года, отсутствие будущей даты;
– перекрёстные проверки дат - минимальный возраст, допустимый диапазон возраста, дата рождения раньше даты выдачи;
– серия и номер паспорта - длина серии, соответствие региона, год печати бланка, длина номера, диапазон номера, редкие номера, согласованность серии и номера, соответствие серии и региона подразделения;
– код подразделения - формат, наличие в реестре, третья цифра;
– машиночитаемая зона - валидация при наличии данных MRZ.

Статусы проверок: passed - проверка пройдена; error - проверка не пройдена (поле message содержит описание ошибки); skipped - проверка пропущена при недостаточности данных.

Программный интерфейс (REST API)

Система предоставляет REST API на основе фреймворка FastAPI со следующими возможностями:

– приём файлов документов на обработку;
– асинхронная обработка с постановкой задач в очередь;
– получение статуса и результатов обработки по идентификатору задачи;
– возврат результатов в формате JSON;
– автоматически генерируемая документация API в формате OpenAPI (Swagger UI, ReDoc);
– функционирование без доступа к внешним сетям.

Перечень эндпоинтов REST API:

Метод	Путь	Описание
POST	/api/tasks	Создание задачи распознавания
GET	/api/tasks/{id}	Статус и результаты задачи распознавания
POST	/api/check	Создание задачи проверки качества изображения
GET	/api/check/{id}	Статус и результаты проверки качества
POST	/api/verify/passport	Верификация паспортных данных

Метод	Путь	Описание
GET	/api/verify/{id}	Результат верификации
GET	/api/health	Состояние системы и метрики
GET	/api/stats	Статистика обработки
GET	/api/archive	История задач
GET	/api/config	Конфигурация для клиента

Режимы функционирования

Система функционирует в следующих режимах:

– штатный режим - непрерывная обработка запросов через REST API в автоматизированном режиме без участия оператора;
– автономный режим - функционирование в изолированной локальной сети без доступа к сети Интернет.

4. АРХИТЕКТУРА

При разработке и эксплуатации программного обеспечения используется следующий технологический стек:

Категория	Технология	Назначение
Язык программирования	Python	Основной язык разработки
Веб-фреймворк	FastAPI	REST API и веб-интерфейс
Фреймворк машинного обучения	PyTorch	Инференс нейронных сетей
Компьютерное зрение	OpenCV	Обработка изображений
Детекция объектов	Ultralytics YOLO	Детекция и сегментация документов
Обработка PDF	PyMuPDF	Извлечение изображений из PDF
Валидация данных	Pydantic	Схемы данных и валидация API
База данных	SQLite	Хранение результатов обработки задач
Контейнеризация	Docker, Docker Compose	Развёртывание и оркестрация

Программное обеспечение реализует слоистую архитектуру:

Слой	Функции
Интерфейсный слой	REST API (FastAPI), документация Swagger UI / ReDoc. Обеспечивает приём запросов и представление результатов.
Слой бизнес-логики	Менеджер задач, фоновый обработчик, управление очередью. Координирует жизненный цикл задач обработки.
Слой конвейера обработки	Экстрактор изображений из файлов, процессор документов. Координирует последовательность этапов обработки.
Слой машинного обучения	Детекторы объектов, классификаторы, модели оптического распознавания символов. Выполняет анализ изображений и извлечение данных на основе нейросетевых моделей.
Слой постобработки	Очистка полей, корректор дат, валидаторы данных. Нормализует распознанные данные и выполняет проверки корректности.
Слой персистентности	База данных SQLite. Хранит результаты обработки задач.

Конвейер обработки документа: входной файл → извлечение изображений → детекция области документа → классификация типа → ротация → детекция информационных полей → оптическое распознавание → постобработка данных → структурированный результат.

Нейросетевые модели системы:

Модель	Архитектура	Назначение
Детектор областей документов	YOLO-seg	Сегментация документа на изображении
Классификатор типов документов	EdgeNeXt + SDTA	Определение типа документа (до 39 классов)
Ротатор ориентации	EdgeNeXt + SDTA	Определение ориентации (0°/90°/180°/270°)
Детекторы информационных полей	YOLO-det	Локализация полей документа
OCR кириллицы	ViTOCR	Распознавание печатного и рукописного текста
OCR машиночитаемых зон	ViTOCR	Распознавание символов MRZ

Все модели оптимизированы для инференса с использованием TorchScript JIT-компиляции и развёрнуты в едином Docker-контейнере.

5. ЮРИДИЧЕСКАЯ ИНФОРМАЦИЯ

Авторские права

Материалы, приведённые в настоящем документе, являются собственностью ООО «Амфител Плюс» и могут быть использованы только специалистами для целей экспертной проверки системы в рамках процедуры включения в Единый реестр российских программ для электронных вычислительных машин и баз данных, а также для личных целей приобретателей программного обеспечения.

Запрещается воспроизведение отдельных частей документа, внесение правок в него, размещение на сетевых ресурсах, распространение в любой форме (в том числе в переводе) на бумажных и электронных носителях, посредством каналов связи и средств массовой информации или каким-либо другим способом без специального письменного разрешения ООО «Амфител Плюс» и ссылки на источник.

Программное обеспечение и товарные знаки, указанные в настоящем документе, принадлежат ООО «Амфител Плюс» и охраняются законом.

Содержание документа

Содержание данного документа может изменяться без предварительного уведомления. ООО «Амфител Плюс» не несёт ответственности за неточности и/или ошибки, допущенные в данном документе, и возможный ущерб, связанный с этим.