Исполнительное резюме “Программный шлюз” для “Уральский Завод Противогололедных Материалов”

Дата создания: 28.07.2025

Наименование документа	Исполнительное резюме
Версия (релиз) для реализации	1
Статус документа	Утверждение
Автор документа	DIAL4AI
Обсуждение и дополнение	Устав, содержание и ограничения проекта могут изменяться по мере планирования. Порядок и содержание изменений определяются

Глоссарий

- **Интеллектуальный интеграционный шлюз** — программный слой между внешним ЭДО (например, Контур.Диадок) и 1С:ERP, который приводит входящие документы к единому стандарту под требования 1С и передаёт их в учётную систему.
- **ЭДО (электронный документооборот)** — внешний канал обмена юридически значимыми электронными документами, из которого поступают файлы для обработки.
- **Контур.Диадок** — провайдер ЭДО, через который принимаются и отправляются электронные документы, доступные для последующей обработки и загрузки в 1С.
- **1С:ERP (1С)** — целевая учётная система, куда попадают проверенные и стандартизированные электронные документы через официальные интерфейсы интеграции.
- **OCR (распознавание текста)** — технология преобразования изображений и сканов в машиночитаемый текст; не применяется к уже «текстовым» PDF/XML.
- **NLP (обработка естественного языка)** — методы интерпретации текста и извлечения реквизитов из неструктурированных документов.
- **LLM (большая языковая модель)** — тип моделей ИИ, применяемых для понимания содержания документа и извлечения недостающих полей, когда формальные шаблоны не подходят.
- **LLM-парсер** — применение LLM для глубокого разбора неформализованных документов с целью выделения ключевых реквизитов.
- **Унифицированное представление документа** — стандартная структура (например, XML/JSON), которую 1С может принять через API без дополнительных преобразований.

- **Интерфейсы 1С** — поддерживаемые способы обмена с 1С (REST API, OData, SOAP), выбор зависит от версии и конфигурации.
- **Форматы данных** — типы файлов, с которыми работает система: PDF, изображения, XML, JSON, PDF/A; по необходимости также CSV (EDI-сообщения) и EML (почтовые вложения).
- **Объектное хранилище S3-совместимого класса** — используемое для хранения временных файлов обработки; совместимо с сервисами, поддерживающими S3-API.
- **SQLite / PostgreSQL** — СУБД для хранения служебных данных: конфигураций, логов и метаданных.
- **On-premise развёртывание** — установка и эксплуатация решения в инфраструктуре заказчика, без привязки к проприетарной облачной платформе.

1. Описание проекта

1.1 Интеллектуальный интеграционный шлюз

Проект предлагает разработку программного шлюза, основанного на технологиях искусственного интеллекта, который станет прослойкой между внешней системой ЭДО (Контур.Диадок) и 1С (1С:ERP). Шлюз автоматически обрабатывает входящие файлы документов, выполняя их интеллектуальный разбор и приведение к единому стандарту. Решение способно распознавать содержимое документов различных типов, классифицировать их, стандартизировать под требования 1С и загружать в систему.

1.2 Ключевые интеллектуальные функции

1.2.1 Обработка и конвертирование

a) Входящие файлы (PDF или изображения) конвертируются в машиночитаемый текст при помощи OCR-модуля
b) Извлеченные данные нормализуются под форматы, поддерживаемые 1С

1.2.2 NLP & LLM

Для понимания неструктурированных фрагментов текста и извлечения ключевых реквизитов используются технологии обработки естественного языка (NLP) и большие языковые модели (LLM). Необходимы для гибкой работы с документами в условиях отсутствия шаблонизированного входного документа.

Шлюз автоматически разделяет пакетные файлы на отдельные документы, преобразует их содержимое в унифицированный электронный формат и передает в 1С через API-интерфейсы. Документы принимаются, распознаются, дробятся на части и регистрируются в 1С автоматически. (см. Приложение 1)

2. Архитектура системы

Архитектура решения имеет модульную структуру. Каждый модуль отвечает за свой этап обработки документа.

2.1 Модуль мониторинга

Отслеживает поступление новых документов из заданных источников (входящие сообщения ЭДО, сетевые папки, email и т.п.) и инициирует процесс обработки.

2.2 Модуль разделения

Выделяет отдельные документы из многостраничных или пакетных файлов (например, разбивает единый PDF с несколькими накладными на индивидуальные файлы по каждой накладной).

2.3 OCR-модуль

Преобразует сканированные страницы и изображения в текст, извлекая машиночитаемые данные из отсканированных PDF или фото-документов.

2.4 LLM-парсер

При необходимости задействует AI-модели для глубокого разбора содержимого неформализованных документов, извлекая ключевые поля и детали из свободного текста. Этот модуль включается только для сложных случаев, экономя ресурсы при обработке типовых документов.

2.5 Генератор

Формирует из извлечённых данных стандартизированный электронный документ или структуру данных, соответствующую форматам и требованиям 1С. Приводит разные входящие документы к единому шаблону, пригодному для импорта в систему.

2.6 Модуль валидации

Проверяет сгенерированный документ на полноту и корректность – контролирует наличие всех обязательных реквизитов, правильность данных, соответствие требуемым форматам (например, XML-схемам) и бизнес-правилам компании. Доступна возможность ручной или частично ручной проверки в данном модуле.

2.9 API-интегратор с 1С

Обеспечивает загрузку проверенных документов непосредственно в 1С через интерфейсы интеграции (REST API, веб-сервисы или внешние обработки 1С). Данный модуль автоматически создаёт в 1С нужные объекты документов (или прикрепляет файлы к уже существующим объектам) согласно результатам обработки.

Примечание:

При масштабировании продукта может быть добавлен модуль классификации для разветвления логики работы с документами иных типов.

3. Принципы работы системы

3.1 Мониторинг поступлений

Модуль автоматически обнаруживает новый документ, поступивший из внешней системы ЭДО (например, полученное через Диадок входящее сообщение с файлом, либо сохранённый файл в оговоренном сетевом каталоге). Обнаружив новый файл, система регистрирует его для последующей обработки, присваивая уникальный идентификатор потока обработки.

3.2 Предварительная обработка и разбиение

Если файл содержит несколько документов (например, пакет сканов накладных в одном PDF), модуль автоматически разделит его на отдельные файлы по

каждому документу. Каждому выделенному документу присваиваются атрибуты (например, порядковый номер, имя файла), и они ставятся в очередь дальнейшего распознавания.

3.3 OCR-распознавание

Для каждого файла, представляющего скан или изображение, модуль OCR выполняет распознавание текста. На этом этапе сканированное изображение документа преобразуется в текстовый слой. Если документ изначально в электронном текстовом формате (например, PDF с текстовым слоем, XML-файл и т.д.), OCR не требуется – система сразу переходит к следующему шагу.

3.4 Парсинг и извлечение данных

В зависимости от типа документа применяется соответствующая логика разбора. Для формализованных документов используются шаблоны и правила извлечения реквизитов. Для неструктурированных документов модуль LLM-парсер подключает AI-модели для семантического анализа текста – например, большая языковая модель извлекает ключевые поля (номер и дату накладной, отправителя, получателя, сумму и пр.) из произвольного текста. Включается при необходимости.

3.5 Формирование унифицированного представления

На основе извлечённых данных формируется стандартная структура данных документа. Информация приводится к единому шаблону, принятому для импорта в 1С. Это может быть XML-файл определённого формата или JSON-структура с необходимыми полями, который 1С сможет принять через API.

3.6 Валидация

Прежде чем импортировать результирующий документ в 1С, он проверяется на полноту и корректность. Контролируется наличие всех обязательных реквизитов, соответствие определенным в ТЗ правилам и техническим требованиям формата. Некорректные или неполные документы помечаются как ошибочные для последующего анализа или ручной корректировки. Информация сохраняется в логе.

Перед переходом к интеграции в 1С, к результирующему документу добавляется вся необходимая дополнительная информация, обусловленная бизнес-правилами компании-заказчика.

3.7 Интеграция в 1С

На финальном этапе документ передается в систему 1С. Интеграция происходит через официальные интерфейсы – например, с помощью REST API веб-сервиса 1С или посредством вызова внешней обработки.

3.8 Уведомление и ручная проверка (при необходимости)

После успешной загрузки шлюз может регистрировать событие (например, в журнале или уведомлении ответственному сотруднику) о том, что документ обработан и доступен в 1С. В сложных случаях система способна отправлять полученный результат на предварительную проверку специалистом. Сотрудник может просмотреть распознанные данные и при необходимости скорректировать их, после чего подтвердить загрузку в 1С.

4. Форматы входных и выходных данных

4.1 Входные данные:

a) PDF-файлы
b) Изображения
c) Структурированные электронные форматы (XML, JSON, PDF/A и др.)
d) Прочие источники - при необходимости можно рассмотреть возможность приёмки CSV(EDI-сообщения), EML(электронная почта)

4.2 Выходные данные:

a) Стандартизированная структура документа для 1С
b) Образ документа (вложение): оригинальный файл документа (или его сформированный PDF, если был получен в ином формате) прикрепляется к соответствующему объекту в 1С
c) Лог и отчёт об обработке

5. Потенциальные технологии для стека и лицензирование

5.1 Потенциальный стек

Язык программирования

- Python(с использованием FastAPI/Django) или C#

OCR и компьютерное зрение

- Open-source OCR: Tesseract OCR
- Предобработка изображений: OpenCV

- Облачные OCR (опционально): Google Cloud Vision, ABBYY Cloud OCR SDK и аналоги

NLP и ML

- Библиотеки/фреймворки: scikit-learn, TensorFlow, PyTorch
- Предобученные модели: BERT/RoBERTa (в т.ч. русскоязычные, RuBERT и др.) или аналоги
- LLM-парсинг: локальный деплой или вызов через API (например, OpenAI API или аналоги)

Интеграция с 1С и внешними системами

- Шлюз к 1С через опубликованные HTTP-сервисы: REST API, OData, SOAP (в зависимости от версии 1С)
- 1С:ERP: публикация внешних интерфейсов / внешние обработки
- Интеграция с Контур.Диадок через их API

Хранение данных

- Метаданные, логи, конфигурации: SQLite / PostgreSQL
- Временные файлы: объектное хранилище (S3-совместимое и т.п.)
- Система учёта результатов: 1С как основное хранилище обработанных документов

5.2 Лицензирование и стоимость владения

При разработке решения упор делается на компоненты с открытым исходным кодом и свободными лицензиями. Следовательно, не потребуется приобретать дорогие лицензии на специализированные продукты распознавания – используются общедоступные библиотечные решения. Исключение могут составлять лишь те части, где без коммерческого продукта не обойтись

(например, если будет принято решение использовать коммерчески лицензируемый модуль ABBYY для повышения точности OCR, или платный сервис LLM).

При необходимости использования облачных модулей (OCR, LLM или хранение данных) выбирается модель с оплатой за фактическое потребление ресурсов.

Архитектура решения не привязана к конкретным проприетарным платформам – при желании шлюз может быть развёрнут в инфраструктуре компании-заказчика (on-premise), без зависимости от облака.

6. Приложения

Приложение 1 (прикладывается к данному документу)

Приложение 2 (прикладывается к данному документу)