|  |  |
| --- | --- |
|  | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

|  |  |
| --- | --- |
| ФАКУЛЬТЕТ | *Робототехника и комплексная автоматизация* |
|  |  |
| КАФЕДРА | *Системы автоматизированного проектирования (РК-6)* |

***ОТЧЕТ ПО*** ***ПРОИЗВОДСТВЕННОЙ ПРАКТИКЕ***

|  |  |  |
| --- | --- | --- |
| Студент |  | Больных Андрей Сергеевич |
|  |  | (фамилия, имя, отчество) |
| Группа |  | РК6-74Б |
| Тип практики |  | Эксплуатационная |
| Название предприятия |  | ООО "ДАТАМАЙНИНГ СОЛЮШИНС" |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Студент | РК6-74Б | |  |  | Больных А. С. |
|  | (Группа) | |  | (подпись, дата) |  |
|  | |  |  |  |  |
| Руководитель практики от кафедры | | |  |  | Оглоблин Д. И. |
|  | |  |  | (подпись, дата) |  |
|  | |  |  |  |  |
|  | |  |  |  |  |
| Оценка |  | |  |  |  |
|  |  | |  |  |  |

*Москва, 2023 г.*

**Министерство науки и высшего образования Российской Федерации**

**Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)» (МГТУ им. Н.Э. Баумана)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | УТВЕРЖДАЮ | | |
|  |  | Заведующий кафедрой | | *РК-6* |
|  |  |  | | (Индекс) |
|  |  |  | |  |
|  |  |  | *А.П. Карпенко* | |
|  |  |  | (И.О.Фамилия) | |
|  |  |  |  | |
|  |  | «\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_ *2023* г. | | |

**ЗАДАНИЕ**

**на прохождение эксплуатационной практики**

|  |  |
| --- | --- |
| Студент | *Больных Андрей Сергеевич, 4* курса, группы *РК6-74Б* |
|  | (фамилия, имя, отчество, № курса, индекс группы) |
|  |  |
| в период с | *«03» июля 2023 г.* по *«30» июля 2023 г.* |
|  |  |
| Предприятие: | ООО "ДАТАМАЙНИНГ СОЛЮШИНС" |
|  |  |
| Подразделение: | Дирекция цифровизации бизнес-процессов |
|  | (отдел/сектор/цех) |
|  |  |
| Руководитель практики от предприятия (наставник): | |
| Журиков Максим Юрьевич, ведущий пресейл менеджер | |
| (фамилия имя отчество полностью, должность) | |
|  | |
| Руководитель практики от кафедры: | |
| Оглоблин Дмитрий Игоревич, руководитель практики | |
| (фамилия имя отчество полностью, должность) | |

Задание.

1. Разработать гибкое описание и определение документа для счёта-фактуры
2. Обеспечить импорт изображений из почты Microsoft Outlook
3. Настроить автоматическое удаление пустых листов в многостраничных документах
4. Реализовать экспорт данных из документов в базу данных MS SQL

Дата выдачи задания *«03» июля 2023 г.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Руководитель практики от предприятия** | | | |  |  | М.Ю. Журиков |
|  | |  | |  | (подпись, дата) |  |
| **Руководитель практики от кафедры** | | | |  |  | Д.И. Оглоблин |
|  | |  | |  | (подпись, дата) |  |
|  | |  | |  |  |  |
| **Студент** | РК6-74Б | |  |  |  | А.С. Больных |
|  | (группа) | |  |  | (подпись, дата) |  |

**СОДЕРЖАНИЕ**

[**ВВЕДЕНИЕ** 4](#_Toc131284932)

[**1.** **ПОСТАНОВКА ЗАДАЧИ** 4](#_Toc131284933)

[**2.** **КРАТКИЙ ОТЧЕТ О ВЫПОЛНЕННЫХ РАБОТАХ** 4](#_Toc131284934)

[**3.** **АНАЛИЗ РЕЗУЛЬТАТОВ** 4](#_Toc131284935)

[**ЗАКЛЮЧЕНИЕ** 5](#_Toc131284936)

[**СПИСОК ЛИТЕРАТУРЫ** 5](#_Toc131284937)

**ВВЕДЕНИЕ**

Документы и прочая корпоративная информация представляют собой немалую ценность так же, как и материальные ресурсы. Чтобы компания была конкурентоспособной, ей нужны современные подходы к обработке документов, при этом важно уделять достаточное внимание вопросам информационной безопасности.

Можно выделить ряд проблем, общих для тех, организаций, где еще активно используют бумагу:

1. документы постоянно теряются;
2. накапливается множество документов, назначение и источник которых не ясны;
3. часто нарушается конфиденциальность – информация попадает в чужие руки;
4. тратится масса времени на поиск нужного документа и формирование комплекта;
5. создается несколько копий одного и того же документа – на бумагу и копирование тратится немало средств;
6. подготовка и согласование документов отнимают много времени.

Внедрение системы автоматизации электронного документооборота позволяет решить все эти проблемы, при этом организация получается следующие преимущества:

* обеспечивается слаженная работа всех подразделений;
* повышается производительность труда сотрудников за счет сокращения времени на создание, обработку и поиск документов;
* повышается оперативность доступа к информации;
* снижаются издержки на обработку документов;
* сокращается время принятия решений.

Автоматизация документооборота – не просто средство оптимизации внутренних процессов, это возможность получить преимущество в условиях жесткой конкуренции. Системы автоматизированного документооборота помогают ускорить процессы, дают руководителям цельную картину по работам компании для принятия как оперативных, так и стратегических решений. Она повышает эффективность процессов любой организации, независимо от масштаба и типа собственности.

Автоматизированные системы документооборота используют технологию OCR (англ. optical character recognition). Оптическое распознавание символов – это процесс преобразования изображения текста в машиночитаемый текстовый формат. Например, при сканировании бланка или квитанции, компьютер сохраняет скан в виде файла изображения. OCR помогает преобразовать изображение в текстовый документ, содержимое которого хранится в виде текстовых данных.

В рамках прохождения эксплуатационной практики **целью** работы является погружение в предметную область документооборота, изучение принципов работы с технологией OCR на примере системы ContentCapture, освоение программного комплекса программ ContentCapture, разработка проекта для автоматизированного документооборота, интеграция различных сторонних сервисов и информационных систем с ContentCapture.

1. **ПОСТАНОВКА ЗАДАЧИ**

На основе OCR системы ContentCapture необходимо разработать проект для автоматизации обработки документов первичной бухгалтерии. Проект будет включать в себя следующие этапы обработки: импорт, классификация, распознавание, извлечение данных, верификация и экспорт.

**Этап 1: Импорт**

Для этапа импорта предусмотрено создание следующих функциональных объектов:

1. **Станция сканирования:** Необходимо настроить и обеспечить работу станции сканирования для получения отсканированных документов.
2. **Сетевая папка на сервере и электронная почта:** Создать сетевую папку на сервере и настроить электронную почту для загрузки отсканированных файлов в проект. Это обеспечит удобный способ передачи документов в систему.
3. **Консольное приложение на языке C#:** Разработать консольное приложение на языке C#, которое будет подключаться к API ContentCapture и осуществлять загрузку пакета документов.

**Этап 2: Классификация**

На этапе классификации требуется разработать и подключить классификатор типов документов. Это позволит системе определить тип каждого документа из списка, указанного в техническом задании (счёт-фактура, УПД, товарная накладная, акт об оказанных услугах).

**Этап 3: Распознавание и извлечение данных**

Для этапов распознавания и извлечения данных необходимо разработать гибкие описания для четырёх видов документов, перечисленных в техническом задании. Это включает в себя следующие шаги:

1. **Распознавание:** Настроить систему ContentCapture на распознавание текста и данных на отсканированных документах.
2. **Извлечение данных:** Разработать механизмы для извлечения необходимых данных с каждого типа документа.
3. **Тестирование шаблонов:** Провести тестирование разработанных шаблонов на небольшом наборе отсканированных файлов, чтобы убедиться в правильности извлечения данных.

**Этап 4: Экспорт данных**

На этапе экспорта данных необходимо создать профили экспорта результатов:

1. **Запись в базу данных MS SQL:** Реализовать профиль, который будет записывать извлеченные атрибуты ключевых полей документов в базу данных MS SQL. Это обеспечит сохранение данных для дальнейшей обработки.
2. **Файл в формате JSON:** Создать профиль экспорта данных в формате JSON, включая извлеченные атрибуты и исходники изображений страниц в формате base64. Это позволит сохранить данные в удобном для долгосрочного хранения и анализа формате.

**Дополнительные задачи**

Дополнительно, следует выполнить следующие задачи:

1. **Удаление пустых листов:** Настроить систему на автоматическое удаление пустых листов в многостраничных документах, чтобы улучшить эффективность обработки.
2. **Автоматическая ориентация страниц:** Разработать механизм для автоматической ориентации страниц документов, что позволит корректно обрабатывать документы с разной ориентацией страниц.
3. **Журнал ошибок и логов:** Создать специальный журнал ошибок и логов для каждого необработанного отсканированного файла. Это поможет в отслеживании и устранении возможных проблем в процессе обработки.

В результате выполнения этого проекта, компания получит эффективную систему для автоматизации обработки документов первичной бухгалтерии, что позволит сэкономить время и ресурсы при ведении бухгалтерии.

1. **КРАТКИЙ ОТЧЕТ О ВЫПОЛНЕННЫХ РАБОТАХ**
   1. **Разработка гибкого описания для счёта-фактуры**

Гибкое описание в AI Layout программного комплекса ContentCapture – это формализованное описание структуры документа. Документом считается совокупность одного или нескольких изображений страниц. Гибкое описание состоит из двух типов объектов: элементы (регион поиска блоков) и блоки (поля с информацией). Цель создания гибкого описания – нахождение блоков. Различают несколько видов документов, а именно:

* Структурированные документы: анкеты, листы переписи населения, налоговые декларации, экзаменационные тесты (ключевые поля, слова находятся в строго определённых местах);
* Слабоструктурированные документы: заявления физических и юридических лиц, запросы на предоставление сведений и копий документов (ключевые поля, слова находятся в разных местах);
* Неструктурированные документы: письма, обращения граждан (ключевые поля, слова могут как быть, так и нет);

Для каждого вида документов поиск информации реализуется различными комбинациями элементов. Вначале создаются элементы, ограничивающие регионы поиска информации в различных предраспознанных системой частей документа. Ниже представлены некоторые примеры элементов (рис. 1):

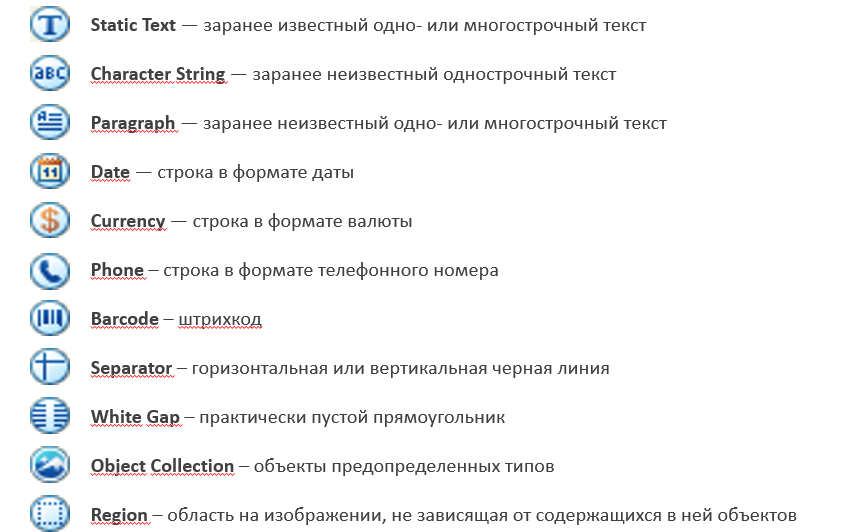
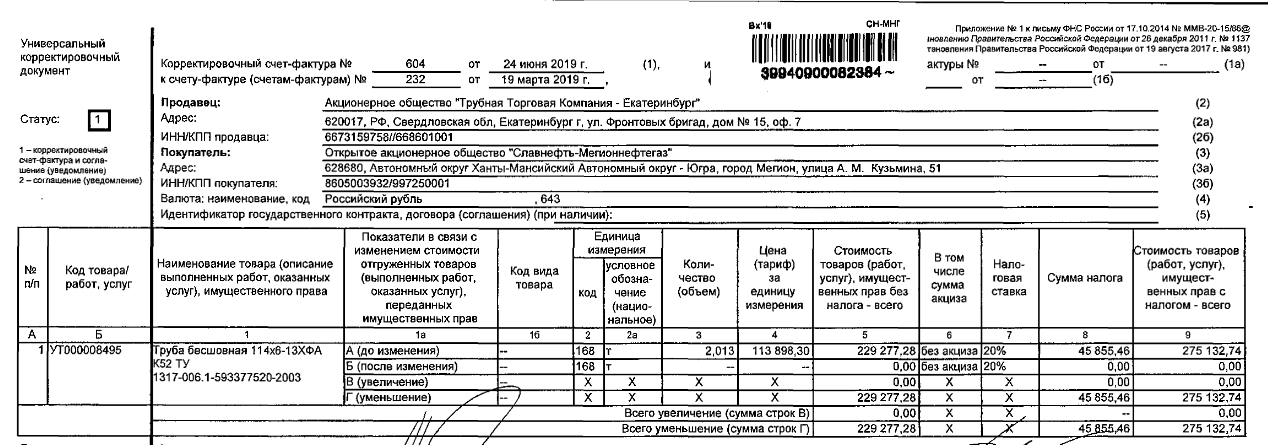
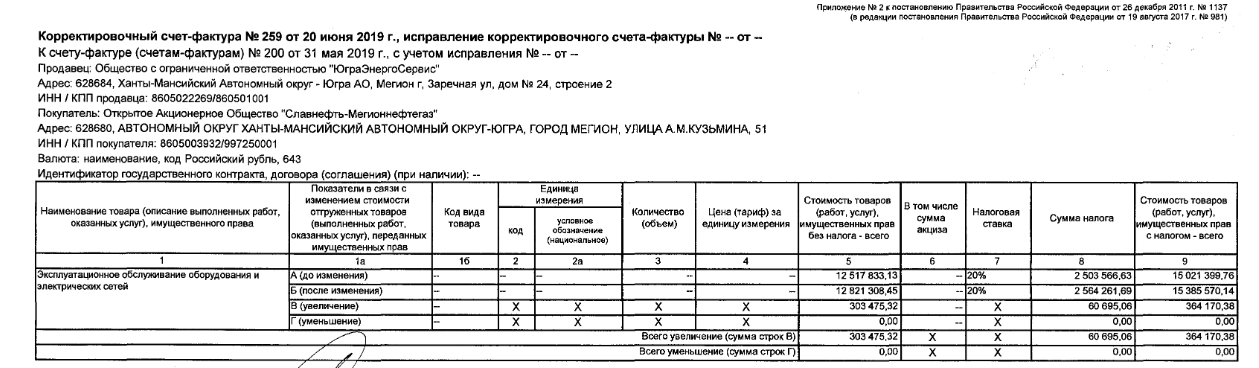


Рисунок 1. Типы элементов гибкого описания

Обычно для поиска нужных полей применяют элемент Static Text -специальное слово или набор символов, координаты которого фиксированы на документе. Это позволит накладывать гибкое описание на разные отсканированные изображения одного типа документа. Например, чтобы система распознала номера счетов-фактур (см. рис. 2а, 2б) создаётся элемент StaticText («якорь»), который ищет ближайшее к верхней границе словосочетание счет-фактура со знаком №. Это позволит ограничить регион поиска нужных цифр.



Рисунки 2а, 2б. Примеры счётов-фактур

После разметки всех регионов и добавления нужных элементов, требуется создать блоки, в которые будет записываться нужная информация.

OCR система ContentCapture работает по принципу оценки гипотез, т.е. разные похожие элементы будут составлять дерево гипотез. Система будет выбирать тот вариант, который соответствует параметрам задания гипотез (уверенность распознавание символа, процентное соотношение ошибочных/лишних символов в словосочетании и т.д.). Чтобы дерево гипотез на одних и тех же элементах не расширялось, нужно группировать элементы в Group.

По окончании тестирования гибкого описания на выборке элементов, файл нужно загрузить в программу Станцию настройки проектов.

* 1. **Настройка импорта изображений через электронную почту**

ContentCapture позволяет импортировать файлы различными способами. Одним из них является создание «горячей» папки на почтовом сервисе, из которого система может в автоматическом режиме забирать информацию из прикреплённых файлов к письмам. Для подключения почтовых сервисов необходимо в Станции настройки проектов создать профиль импорта изображений (см. рис. 3).

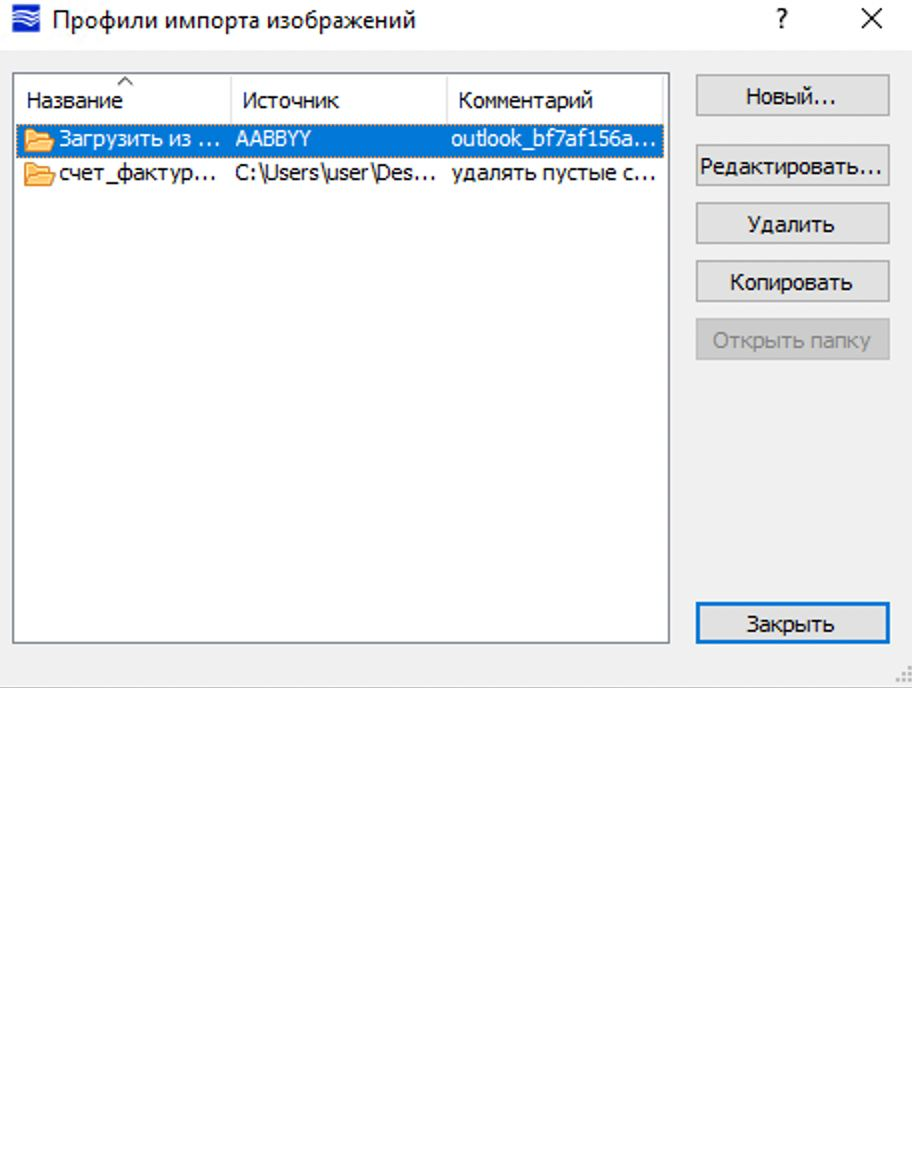


Рисунок 3. Добавление профиля импорта изображений

Для использования сервиса MS Outlook в поле выбора «Горячая папка» необходимо указать Microsoft Graph Mail API (см. рис. 4). Система также поддерживает работу с FTP, IMAP, POP3, SFTP серверами.

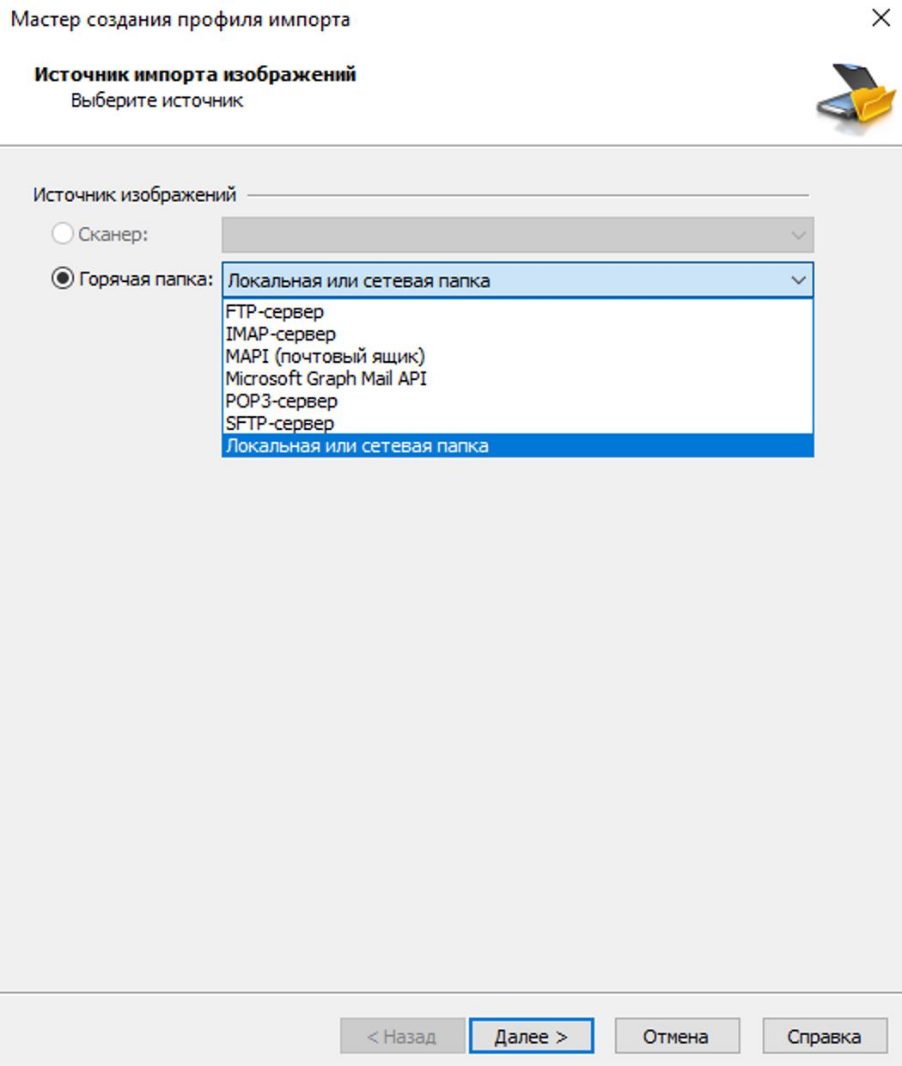


Рисунок 4. Выбор почтового сервиса

В результате создания профиля импорта изображений, проект получил возможность обращаться к почтовой папке для принятия отсканированных документов. На рисунке 5 представлено вложение из Microsoft Outlook, ниже импортированный и распознанный документ в отдельном пакете командой «Импортировать». Стоит отметить, что новые документы могут добавляться в уже созданные пакеты – также есть функция автоматической проверки папки на наличие новых писем (см. рис. 7).

*Примечание: термин пакет означает стопку одновременно отсканированных или импортированных из определённого каталога документов. Это внутренний формат для работы с пересылкой документов на разные этапы обработки.*

К сожалению, в ходе выполнения и тестирования профиля импорта изображений мне не удалось обеспечить непрерывную отправку вложений из писем с помощью данной функции, даже с большим периодом проверки писем (более 2 минут). Возможно, автоматические запросы о доступе к письмам из вложений блокировались сервисами анти-спама MS Outlook.

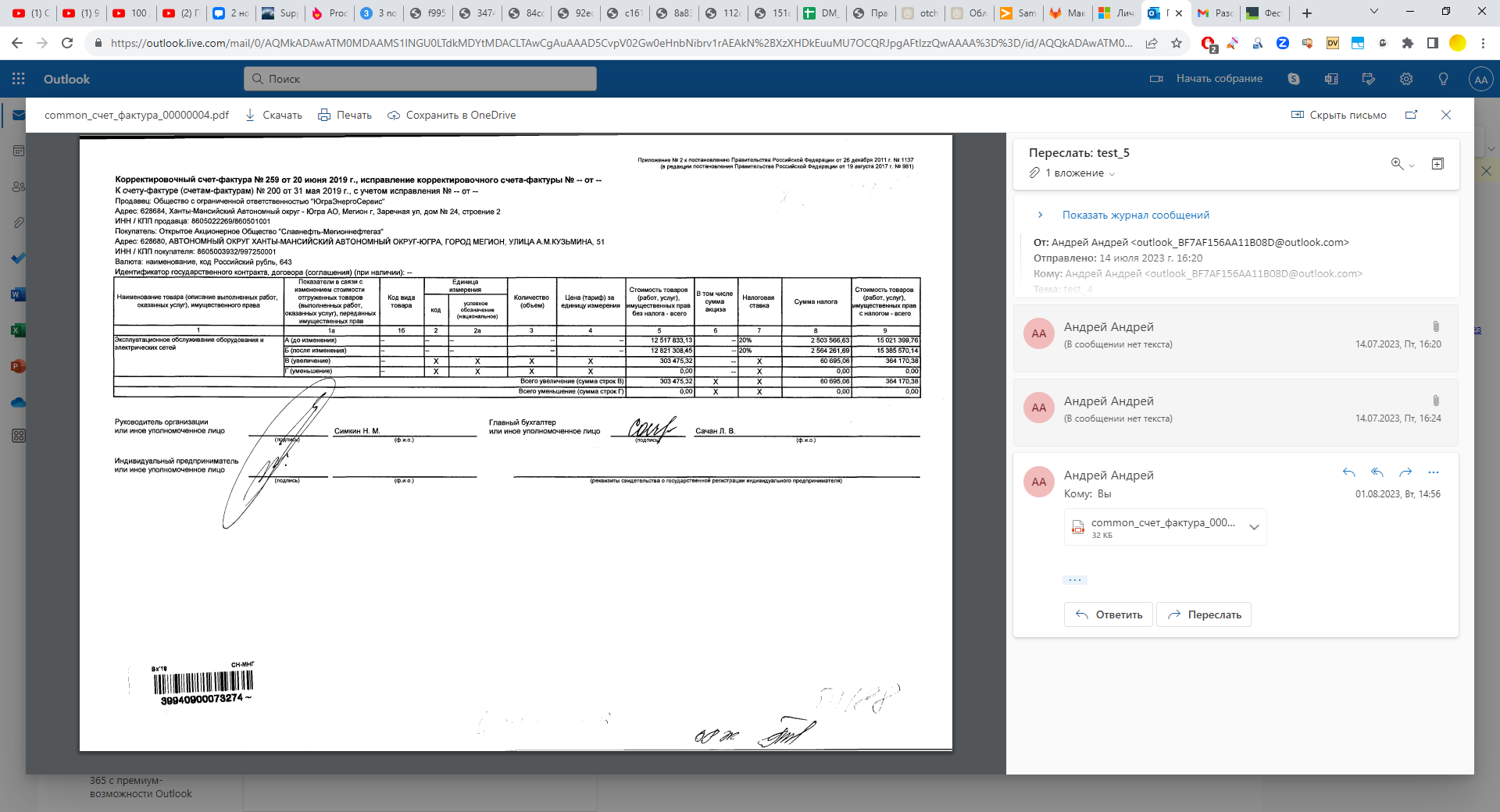


Рисунок 5. Пример вложения письма в MS Outlook

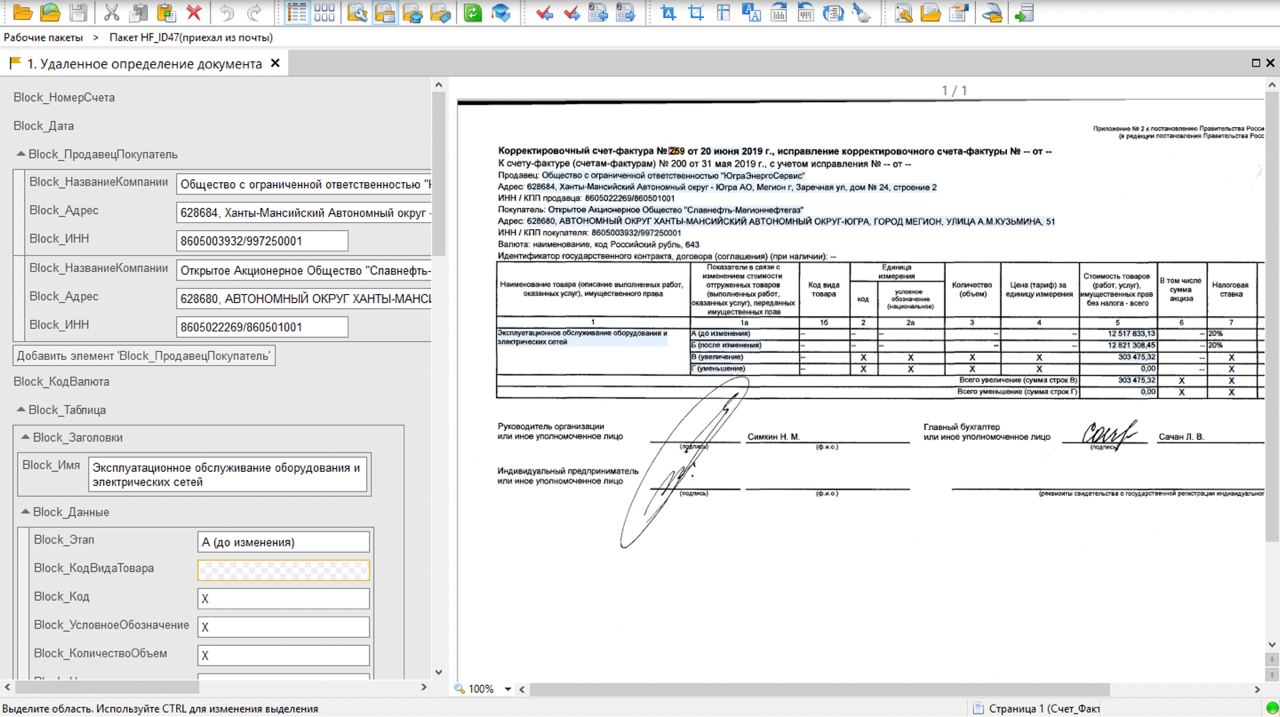


Рисунок 6. Пример документа из MS Outlook в ContentCapture

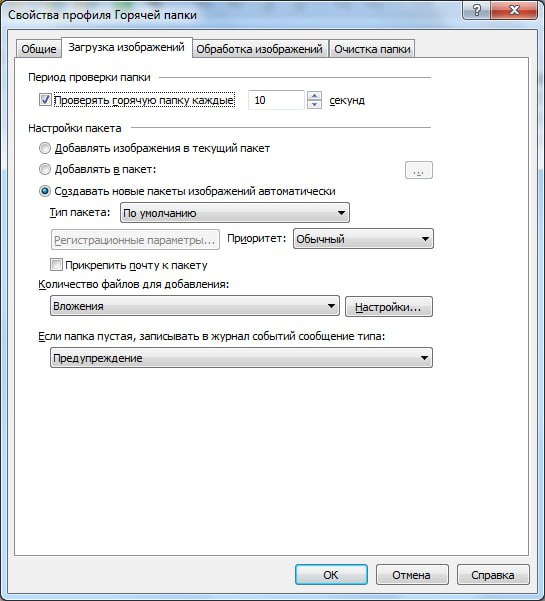


Рисунок 7. Свойства и параметры почтовой папки

* 1. **Экспорт данных в базу данных MS SQL**

Для настройки экспорта данных в базы данных (БД) необходимо перейти в Станцию настройки проектов и вызвать функцию «Мастер настройки приёмника экспорта» (см. рис. 8).

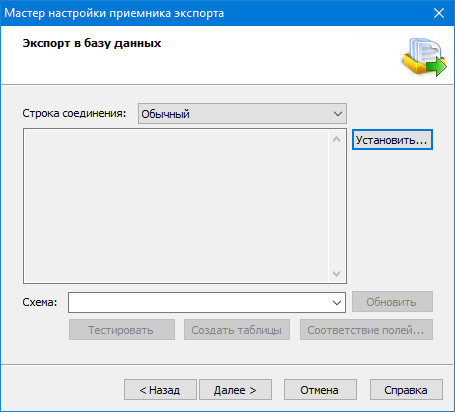


Рисунок 8. Диалоговое окно мастера настройки экспорта в БД

Затем необходимо правильно выбрать драйвер для создания канала передачи данных и работы с СУБД в зависимости от типа и версии БД, в которую будут экспортироваться данные. Для MS SQL был выбран ODBC Driver (см. рис. 9).

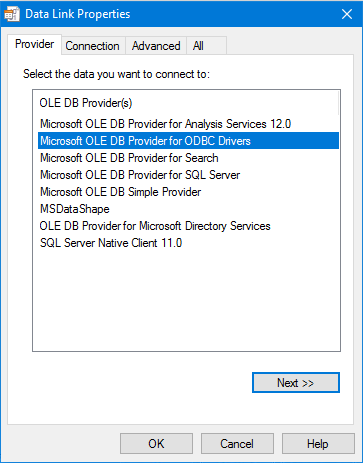


Рисунок 9. Выбор драйвера для работы с БД

После этого создаётся config-файл соединения, в который нужно ввести корректные атрибуты для соединения с БД, такие как: адрес сервера, логин и пароль пользователя сервера БД, рабочая схема БД.

В результате при создании задачи экспорта в Станции настройки проектов, информация из пакета распознанных документов запишется в БД – ключами станут названия блоков, а их значения будут верифицированные значения блоков.

* 1. **Автоматическое удаление пустых листов**

ContentCapture предоставляет широкие функции для предобработки различных типов документы, а именно:

* Вставить страницы
* Повернуть
* Отразить
* Обрезать
* Инвертировать
* Конвертировать в ч/б
* Исправить перекос
* Очистить от мусора
* Скрыть данные
* Перемещать страницы внутри пакета документов

Часть из них настраивается в свойствах проекта, рисунок 10. Функция удаления пустых страниц может быть модифицирована дополнительными настройками, например, удалять страницы только в определённых типах документов. Также предобработку документов можно делать в Станции Сканирования программного комплекса ContentCapture. Оттуда же, можно импортировать профили улучшения изображений (см. рис. 11) – специально настроенные шаблоны для разных видом документов, пример: документ с большим количеством печатей.

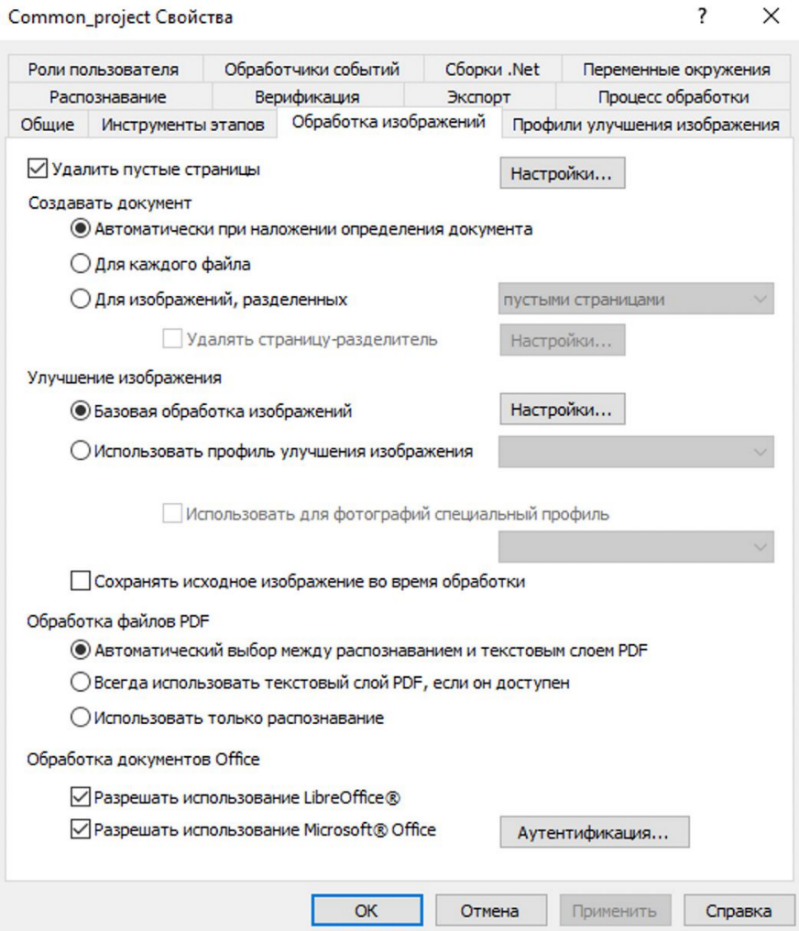


Рисунок 10. Свойства обработки изображений

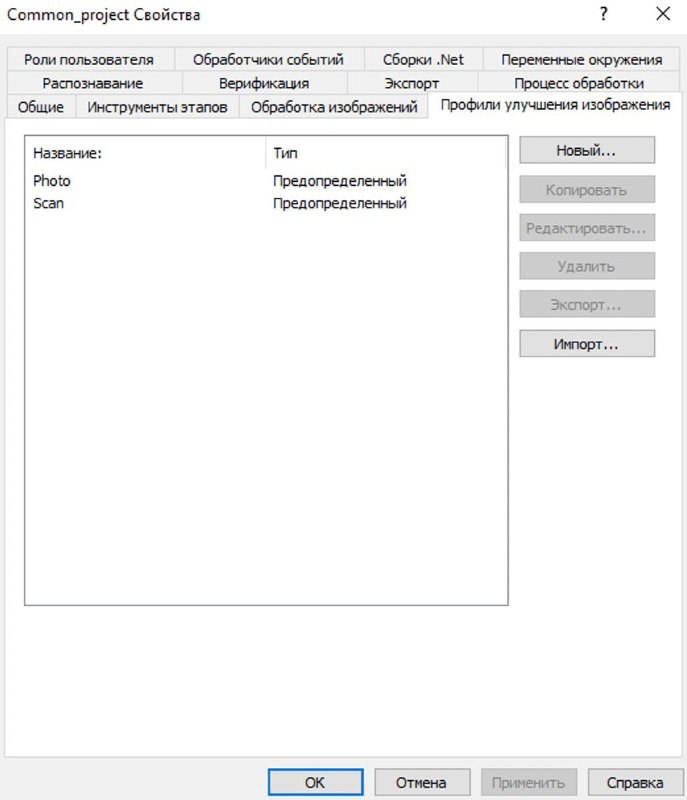


Рисунок 11. Окно профилей улучшения изображений

1. **АНАЛИЗ РЕЗУЛЬТАТОВ**

Функционал для разработки гибкого описания оказался довольно обширным и комплексным, но при этом больше половины встроенных возможностей оказались в нерабочем состоянии и/или они работали некорректно. Стоит отметить, что для работы некоторых узконаправленных функций требовались отсканированные документы с минимальным числом «артефактов» и «шумов», т.е. очень хорошего качества. Встроенные возможности ContentCapture иногда позволяли удалить нежелательные ошибки сканирования и улучшить качество электронных документов, но далеко не всегда. Стоит отметить, что у меня единственного в команде практикантов система в разных программах ContentCapture собственноручно меняла местами информацию из блоков. В AI Layout Studio гибкое описание отработало некорректно с извлечением информации об ИНН клиента и поставщика (см. рис. 6). При последующем импорте файла гибкого описания в Станцию настройки проектов, где происходили все этапы обработки документов, система распознавания текста записывали номера ИНН не сверху-вниз, а снизу-вверх – вопреки всем правилам и «якорям» для поиска регионов текста.

При создании профиля импорта изображений некоторые почтовые сервисы оказались несовместимыми с программным обеспечением ContentCapture, но в целом интеграция почты и её собственного функционала для автоматической передачи сканов изображения на сервер с последующим преобразованием файлов во внутренний формат прошла легко и успешно с минимумом ошибок.

Экспорт данных в базу данных оказался трудоёмкой задачей. Это связано с недостаточно подробным документированием данного функционала. Подробных инструкций по соединению сервера проекта бухгалтерии с сервером СУБД – нет. Данная задача потребовала дополнительного изучения навыков системного и сетевого администрирования. Но в конечном итоге, после настройки подключения с базой данных и СУБД ContentCapture успешно экспортировала данные из документов в автоматическом режиме.

**ЗАКЛЮЧЕНИЕ**

В ходе прохождения эксплуатационной практики в компании DM Solutions я принял участие в командной разработке серверного проекта для автоматизации потокового документооборота. Мною были изучены особенности и аспекты электронной бухгалтерии, различные виды типовых документов – их назначения и отличия. Я приобрёл навыки работы с lowcode студией AI Layout для создания описания шаблона документа, которое впоследствии можно улучшить благодаря дообучению программы-распознавателя.

Также я поработал с различными API для автоматического импорта отсканированных документов и экспорта данных из них. Программное обеспечение ContentCapture, используемое компанией DM Solutions, продемонстрировало промышленные масштабы различных встроенных функций и широкие возможности для интеграции с различными ERP, CRM, ECM, СЭД, АБС и другими информационными системами. Стоит отметить, что в течение работы с программным комплексом ContentCapture, данное программное обеспечение показало крайне низкое число недостатков при реализации некоторых специфических задач.

Таким образом, российское программное обеспечение ContentCapture является универсальным, легко масштабируемым и гибким решением для автоматизированного документооборота корпоративных предприятий.

**СПИСОК ЛИТЕРАТУРЫ**

1. Автоматизация документооборота. Актуальные вопросы и ответы [Электронный ресурс] – URL: <https://www.directum.ru/blog-post/avtomatizacija_dokumentooborota_v_voprosakh_i_otvetakh> (дата обращения: 25.07.2023)
2. ABBYY FlexiCapture [Электронный ресурс] – URL: https://www.id-mt.ru/produkty/abbyy/abbyy\_flexicapture/ (дата обращения: 12.07.2023)
3. Универсальная платформа для интеллектуальной обработки информации [Электронный ресурс] – URL: https://contentai.ru/datacapture (дата обращения: 25.07.2023)
4. ContentCapture Distributed [Электронный ресурс] – URL: <https://contentai.ru/contentcapture-downloads/distributed/> (дата обращения: 05.07.2023)
5. Учебные материалы по ABBYY FlexiCapture [Электронный ресурс] – URL: https://dnstuff.ru/sharing/elEudrvia (дата обращения: 04.07.2023)
6. FlexiCapture 12 System Administrator's Guide [Электронный ресурс] – URL: https://help.abbyy.com/en-us/flexicapture/12/admin\_guide/introduction\_ag/ (дата обращения: 05.07.2023)