Руководство инженера.

Система SOICA.

Версия 4.0.2

Утверждено:

Горюнов В. Н.

Подпись:

Дата редактирования:

Создано:

Гончаров Г. А.

Польшаков В. С.

Дата: 28.05.2019

29.06.2023

Оглавление

Определения:

ТЗ (Техническое задание) - документ, содержащий требования заказчика.

Границы Кенни – грани изображения на документе найденные по алгоритму Кенни *(Сглаживание, Поиск градиентов, Подавление немаксимумов, Двойная пороговая фильтрация, Трассировка области неоднозначности).* Подробнее: *https://ru.wikipedia.org/wiki/%D0%9E%D0%BF%D0%B5%D1%80%D0%B0%D1%82%D0%BE%D1%80_%D0%9A%D1%8D%D0%BD%D0%BD%D0%B8*

Метод Хафа – метод использующий идентификацию прямых в изображении, а также идентификацию позиции произвольной фигуры, чаще всего эллипсов и окружностей. Подробнее: https://ru.wikipedia.org/wiki/%D0%9F%D1%80%D0%B5%D0%BE%D0%B1%D1%80%D0%B0%D0%B7%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5_%D0%A5%D0%B0%D1%84%D0%B0

Каскад Хаара – набор признаков Хаара. Признаки Хаара — признаки цифрового изображения, используемые в распознавании образов. Подробнее: https://ru.wikipedia.org/wiki/%D0%9F%D1%80%D0%B8%D0%B7%D0%BD%D0%B0%D0%BA%D0%B8_%D0%A5%D0%B0%D0%B0%D1%80%D0%B0

ДУЛ – документ удостоверяющий личность.

REST API – способ взаимодействия систем используя архитектурный стиль компонентов REST. https://ru.wikipedia.org/wiki/REST

Документ – классифицированная страница (или набор страниц). У каждого документа свои найденные данные, на валидации своя карточка документа.

Пакет – набор документов. В пакете могут находится как связанные между собой документы, так и нет. Все зависит от настроек импорта файлов в систему.

Профиль очистки – настраиваемый комплекс предобработки страницы (изображения) для улучшения поиска необходимых данных.

Профиль распознавания – совокупность настроек, использующая профили очистки (может быть несколько), движки распознавания текста (ocr), выбор языка, сегментация и т. д.

Класс пакета – отдельный проект в настроечной среде. Сценарий, по которому система будет проводить поступающие с импорта страницы. В классе пакета указываются настройки импорта, условия классификации страниц в документы, профили предобработки и распознавания. Количество классов пакета не ограничено.

Класс документа –форма настраиваемого документа. Относится к конкретному классу пакета. В классе документа указывается какие данные необходимо извлекать, по какой логике извлекаются данные, по какому сценарию экспорта будет экспортирован итоговый документ, форма валидации документа и т. д. Количество классов документа не ограничено.

Локатор – инструмент извлечения данных с документа. Их более 30 видов. Они могут использовать ранее выполненные локаторы. Используемые локаторы называются наследуемыми, а локаторы в которых они используются – зависимыми.

Подполя локатора – атрибуты выполненного локатора. В зависимости от вида локатора у него есть набор подполей (минимум 1 подполе). Например, ключ (слово относительного которого происходил поиск), область (область в которой был произведен поиск) и результат (найденные данные в указанной области).

Альтернатива подполя - результат работы локатора. У каждого подполя может быть несколько альтернатив. Например, под одно регулярное выражение даты может подходить несколько данных с документа, соответственно подполе будет «Результат», а **альтернатив** данного подполя несколько (12.12.2000, 0906.1997).

Поле – конечное значение, передаваемое на валидацию или в экспорт (если валидно). Поле наследует альтернативу подполя какого-либо локатора.

Таблица – конечное значение в виде таблицы. Наследуется только от локаторов с типом «таблица».

Валидность – соответствие найденных данных с данными на изображении.

Степень доверия – вероятность того что найденные системой данные валидны. Измеряется в процентах.

Доксет – набор страниц для настройки проекта. На примере этих страниц настраивается и проверяется корректность сценария обработки и извлечения данных с изображений. Количество страниц в доксете не ограничено.

1. Описание системы.

SOICA – программный продукт по потоковому вводу данных и настройке умного извлечения бизнес-данных из любых носителей текстовой информации.

Основные преимущества системы над конкурентами:

Простота – настройка без кода;

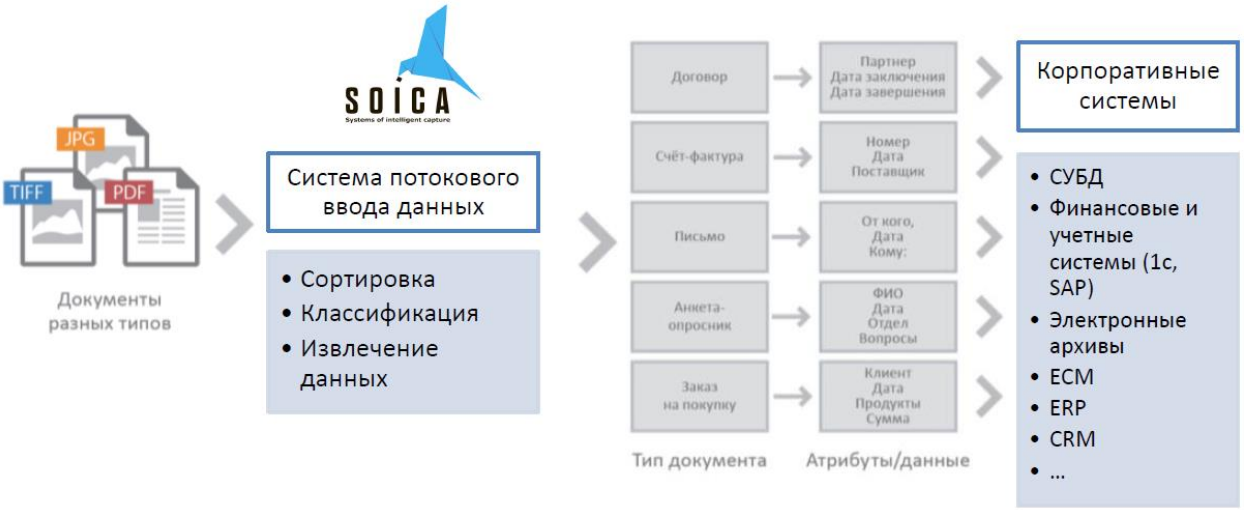
Высокий уровень абстрактности - находятся даже непредсказуемые данные;

Революционная гибкость – сами учим систему думать и принимать решения;

Никаких ограничений - поиск данных на любых носителях текстовой информации.

Предназначение системы:

Система предназначена для потоковой обработки данных. Обработка включает в себя распознавание текста, автоматическую классификацию документа, проверку на валидность, вывод данных в нужном виде, а также загрузки данных в корпоративную систему при необходимости.



После внедрения системы уменьшается время на обработку документов, экономится время и деньги на исправление ошибок, нет необходимости в большей части рутиной работы.

Система направлена на автоматизацию процессов работы с документами в компаниях с документооборотом более 10 000 страниц в год.

Данные для обработки могут быть получены из:

- Каталог на ПК (сетевой каталог);
- Электронная почта;
- FTP;
- Внешние системы (через REST API);
- Мобильное приложение.

Импортировать можно как одностраничные, так и многостраничные документы. Поддерживаемые форматы изображений: JPEG, PDF, TIFF, BMP, PNG, DOCX, GIF

В системе предусмотрено многоканальное импортирование и экспортирование. **Что это значит:** По одному и тому же сценарию может одновременно приниматься файлы разными способами и выдавать нужные данные так же в нескольких вариантах одновременно. **Пример:** готовые данные могут одновременно попасть в файл xml в заданную папку и прийти по электронной почте.

Экспорт полученных данных может идти по разным маршрутам:

- Каталог на ПК (сетевой каталог);
- Электронная почта;
- FTP;
- Через сервис REST API;
- Мобильное приложение.

При отправке результатов по электронной почте адресат может быть задан заранее или сформирован на основе данных находящихся на распознанных документах. В качестве адресата может выступать отправитель, если документы пришли в SOICA по электронной почте.

В портфель решений «SOICA» входит набор продуктов для успешного решения любых задач по оцифровке документов, их классификации и разделению по заранее настроенным сценариям.

Адаптивная корректировка качества сканируемых документов позволяет очистить документы от мусора и повысить качество получаемых изображений.

При работе с очищенным изображениями достигается максимально точное нахождение данных используя внутренние инструменты системы.

Система SOICA имеет общий сценарий работы:

1. **Импорт**. В систему загружаются файлы для распознавания и дальнейшей обработки
2. **Очистка**. С помощью профилей очистки идет подготовка изображения для дальнейшего более качественного распознавания.
3. **Распознавание**. Процесс в результате которого создается репрезентация (результат распознавания), содержащая преобразованное изображение и, возможно, результаты OCR. На изображении происходит нахождения всех имеющихся символов и слов для дальнейшей работы с ними.
4. **Классификация**. В системе предусмотрено несколько способов классификации документа. На основе полученного OCR (выборка ключевых слов или фраз), на основе цветовой гаммы изображения, на основе штрих-кода, на основе изображения лица.
5. **Извлечение данных.** С помощью настроенных сценариев извлечения данных (пункт 2.4 Меню поиска данных (Классы документов)) происходит нахождение конкретных значений на изображении. Значения ищутся согласно ТЗ заказчика.
6. **Форматирование и валидация.** После извлечения данных их необходимо привести в вид согласно ТЗ, а также проверить на правильность нахождения.
7. **Вывод результата.** Наглядное выведение найденных данных для их проверки (пункт 4. Модуль Валидации).

8. **Экспорт**. Согласно настроенному сценарию экспорта, найденные и одобренные на этапе валидации данные экспортируются в определенном формате. Формат определяется в ТЗ (пункт 5. Экспорт)

REST API.

В системе предоставлена возможность работы с сервисами REST. С помощью этих сервисов можно загрузить файлы в систему для обработки, так же использовать при экспорте и управлении пакетами.

Аутентификация

Для получения доступа к функциям rest-сервиса требуется аутентификация – BasicHttpAuth, логин и пароль передаются в заголовке http, при этом логин в открытом виде, а пароль – зашифрованный MD5.

Rest-сервис импорта

Базовый адрес: http://{адрес сервера}/SoicaService/AutoImport.svc

1. Создание пакета: /Create/{batchClass}/{name}, где batchClass - наименование класса пакета, name - наименование пакета, картинка передается как поток.
   Http-метод: POST
   Результат: guid пакета
   Клиентский метод на C#:

```
public string CreateBatch(string className, string name, Stream content)
{
        var address = String.Format(@"{0}/AutoImport.svc/Create/{1}/{2}",
        _baseAddress, className, name);
        _logger.Debug("Адрес {0}", address);
        HttpWebRequest request = (HttpWebRequest)HttpWebRequest.Create(address);
        request.Timeout = _timeout;
        request.Method = "POST";
        request.SendChunked = true;
        request.AllowWriteStreamBuffering = false;
        request.ContentType = MediaTypeNames.Application.Octet;
        var authInfo =
Convert.ToBase64String(Encoding.UTF8.GetBytes(String.Join(":", _login_password)));
        request.Headers.Add(HttpRequestHeader.Authorization, "Basic " + authInfo);
        request.Headers.Add("serviceName", "autoimport");
        using (Stream outputStream = request.GetRequestStream())
        {
                content.CopyTo(outputStream);
        }
        var response = (HttpWebResponse)request.GetResponse();
        if (string.IsNullOrEmpty(response.CharacterSet))
                return String.Empty;
        var encoding = Encoding.GetEncoding(response.CharacterSet);
        string result = String.Empty;
        using (var responseStream = response.GetResponseStream())
        using (var reader = new StreamReader(responseStream, encoding))
        result = reader.ReadLine();
```

```
        response.Close();
        return result.Trim('"');
    }
```

2. Создать пакет и назначить его на определенного пользователя/группу: /Create/{batchClass}/{name}/{login}, где batchClass - наименование класса пакета, name - наименование пакета, logn - логин пользователя или группы пользователей, картинка передается как поток.
   Http-метод: POST
   Результат: guid пакета.

3. Добавить файл в пакет: /AddToBatch/{batchGuid}, где batchGuid - guid пакета, картинка передается как поток.
   Возвращает признак успешности операции true или false.
   Http-метод: POST
   Клиентский метод на C#:

```
    public void AddPageToBatch(string batchGuid, Stream content)
{
        HttpWebRequest request =
(HttpWebRequest)HttpWebRequest.Create(String.Format(@"{0}/AutoImport.svc/AddToBat
ch/{1}", _baseAddress, batchGuid));
        request.Timeout = _timeout;
        request.Method = "POST";
        request.SendChunked = true;
        request.AllowWriteStreamBuffering = false;
        request.ContentType = MediaTypeNames.Application.Octet;
        var authInfo =
Convert.ToBase64String(Encoding.UTF8.GetBytes(String.Join(":", _login_password)));
        request.Headers.Add(HttpRequestHeader.Authorization, "Basic " + authInfo);
        request.Headers.Add("serviceName", "autoimport");
        using (Stream outputStream = request.GetRequestStream())
        {
            content.CopyTo(outputStream);
        }
        var response = (HttpWebResponse)request.GetResponse();
        response.Close();
}
```

4. Запустить пакет на обработку: /StartProcess/{batchGuid}
   Http-метод: GET
   Результат: признак успешности операции – true или false.

5. Создание пакета с дополнительными параметрами: /CreateForUserWithParam/{batchClass}/{name}/{login}/{fileId}/{folderId}/{clientId}/{dogovor}/{serviceId}/{ticket}, где:
   batchClass - наименование класса пакета
   name - наименование пакета
   login - имя пользователя/группы
   fileId - идентификатор исходного файла в ECM
   folderId - идентификатор папки, в которой размещен исходный файл в ECM
   clientId - идентификатор клиента
   dogovor - идентификатор договора на аутсорсинг
   serviceId - идентификатор услуги
   ticket - номер запроса

файл передается как поток в теле запроса
Http-метод: POST
Результат: guid созданного пакета

6. Создание пакета с шаблоном docx:
   /CreateTemplate/{batchClass}/{name}, где:
   batchClass - наименование класса пакета
   name - наименование пакета
   файл шаблона docx передается как поток в теле запроса
   Http-метод: POST
   Результат: guid созданного пакета

Rest-сервис экспорта

Базовый адрес: http://{адрес сервера}/SoicaService/RemoteExport.svc

1. Получить имена пакетов, доступных для экспорта: /BatchNames
   Http-метод: GET
2. Получить имена классов пакетов: /BatchClassNames
   Http-метод: GET
3. Получить имена пакетов определенного класса пакета, доступных для экспорта: /BatchNamesByBatchClass/{batchClassName}
   Http-метод: GET
4. Получить guid пакета по классу пакета и наименованию пакета: /BatchGuidByBatchName/{batchClassName}/{batchName}
   Http-метод: GET
5. Получить наименование класса пакета по guid пакета: /BatchClassNameByBatchGuid/{batchGuid}
   Http-метод: GET
6. Получить пакет по guid: /Batch/{batchGuid}
   Http-метод: GET
7. Получить список имен документов по guid пакета: /DocumentNames/{batchGuid}
   Http-метод: GET
8. Получить документ по guid пакета и наименованию документа: /Document/{batchGuid}/{docName}
   Http-метод: GET
9. Получить список имен полей документа по guid Пакета и наименованию документа:/FieldNamesByDocName/{batchGuid}/{docName}
   Http-метод: GET
10. Получить данные поля по guid пакета, наименованию документа и наименованию поля: /Field/{batchGuid}/{docName}/{fieldName}
    Http-метод: GET
11. Получить список имен таблиц в документе по guid пакета и наименованию документа: /TableNamesByDocName/{batchGuid}/{docName}
    Http-метод: GET

12. Получить данные таблицы по guid пакета, наименованию документа и наименованию таблицы: /Table/{batchGuid}/{docName}/{tableName}
    Http-метод: GET
13. Получить файл экспортированного пакета (pdf, docx, xlsx, tiff):
    /Image/{batchGuid}/{exportName}
    batchGuid - guid пакета, exportName - наименование настроек экспортирования
    Http-метод: GET
    Пакет предварительно должен быть выгружен модулем экспорта.
    результат: поток данных.
14. Получить файл экспортированного документа (pdf, docx, xlsx, tiff):
    /Image/{batchGuid}/{documentGuid}/{exportName}
    batchGuid - guid пакета, documentGuid - guid документа, exportName - наименование настроек экспортирования
    Http-метод: GET
    Документ предварительно должен быть выгружен модулем экспорта.
    результат: поток данных.
15. Получить файл экспортированного документа (pdf, docx, xlsx, tiff) по имени:
    /Image/{batchGuid}/{documentGuid}/{exportName}/{fileName}
    batchGuid - guid пакета, documentGuid - guid документа, exportName - наименование настроек экспортирования, filename – имя файла
    Http-метод: GET
    Документ предварительно должен быть выгружен модулем экспорта.
    результат: поток данных.
16. Получить список имен файлов экспортированного документа (pdf, docx, xlsx, tiff):
    /ImageNames/{batchGuid}/{documentGuid}/{exportName}
    batchGuid - guid пакета, documentGuid - guid документа, exportName - наименование настроек экспортирования
    Http-метод: GET
    Документ предварительно должен быть выгружен модулем экспорта.
    результат: массив строк.
17. Получить изображения экспортированного пакета (jpg, tiff, png), запакованные в zip-архив: /Images/{batchGuid}/{exportName}
    batchGuid - guid пакета, exportName - наименование настроек экспортирования
    Http-метод: GET
    Пакет предварительно должен быть выгружен модулем экспорта.
    результат: поток данных.
18. Получить изображения экспортированного документа (jpg, tiff, png), запакованные в zip-архив: /Images/{batchGuid}/{documentGuid}/{exportName}
    batchGuid - guid пакета, documentGuid - guid документа, exportName - наименование настроек экспортирования
    Http-метод: GET
    Документ предварительно должен быть выгружен модулем экспорта.
    результат: поток данных.
19. Получить результаты распознавания:
    /ProfileJson/{batchGuid}/{pageNum}/{profile},
    где batchGuid – guid пакета,
    pageNum - номер страницы в пакете,
    profile – наименование профиля распознавания.

Http-метод: GET.
Документ предварительно должен быть выгружен модулем экспорта.
Результат: данные распознавания в формате json:

```
{"Id":Int32, // идентификатор
"Guid":Guid, // guid пакета
"Name":String, // наименование пакета
"PIndex":Int32, // индекс страницы
"PostSkewAngle":Int32, // угол поворота (исправление перекосов)
"RepresentationType":RepresentationTypes, // тип репрезентации
"TextLines":[{
        "IndexPage":Int32, // индекс линии на странице
        "BlockType":PolyBlockType, // тип блока
        "Words":[{
            "IndexPage":Int32, // индекс слова на странице
            "IndexLine":Int32, // индекс слова в линии
        "Language":String, // язык распознавания
        "Angle":Double, // угол поворота
            "WordType":SoicaWordTypes, // тип слова
            "KnownColor":KnownColors, // цвет слова
            "RealH":Int32, // высота слова
            "IsCity":Boolean, // является наименованием города
            "IsName":Boolean, // является именем
            "IsStreet":Boolean, // является наименованием улицы
            "IsCompanyName":Boolean, // наименование компании?
            "IsAbbreviation":Boolean, // аббревиатура
            "IsMail":Boolean, // является Email
            "PIndex":Int32, // индекс станицы
            "Id":Int32, // идентификатор
            "Confidence":Single, // конфиденс слова
            "Text":String, // текст слова
            "Height":Double, // высота
            "Width":Double, // ширина
            "X":Double, // позиция слова по X
            "Y":Double, // позиция слова по Y
            }],
        "Text":String, // текст строки
        "Confidence":Single, // конфиденс
        "X":Double, // позиция строки по X
        "Y":Double, // позиция строки по Y
        "Width":Double, // ширина
        "Height":Double, // высота
        "PIndex":Int32, // индекс строки
        "Id":Int32, // идентификатор строки
    }],
"Blocks":[{ // нетекстовые блоки (подписи, печати, чек-боксы…)
    "X":Int32, // позиция блока по X
    "Y":Int32, // позиция блока по X
    "Width":Int32, // ширина
    "Height":Int32, // высота
    "Confidence":decimal, // конфиденс
    "Text":string, // текст блока (либо название блока)
```

}],
"RecognizeProfileId":Int32} // идентификатор профиля распознавания

20. Получить результаты распознавания:
/ProfileXml/{batchGuid}/{pageNum}/{profile},
где batchGuid – guid пакета,
pageNum - номер страницы в пакете,
profile – наименование профиля распознавания.
Http-метод: GET
Документ предварительно должен быть выгружен модулем экспорта.
Результат: данные распознавания в формате xml:

```
<RecognizeResult xmlns="http://schemas.datacontract.org/2004/07/SoikaAPI.Interfaces"
xmlns:i="http://www.w3.org/2001/XMLSchema-instance">
<Blocks xmlns:a="http://schemas.datacontract.org/2004/07/SOICAII">
        <a:Data i:type="тип дополнительного блока">
                <a:Confidence>конфиденс</a:Confidence>
                <a:Text>текст</a:Text>
                <a:Width>ширина</a:Width>
                <a:Height>высота</a:Height>
                <a:X>x-координата</a:X>
                <a:Y>y-координата</a:Y>
        </a:Data>
</Blocks>
<Guid>guid результата распознавания</Guid>
<Id>идентификатор</Id>
<Img>
        <Name>наименование изображения</Name>
        <Path>путь к изображению</Path>
        <Size
xmlns:a="http://schemas.datacontract.org/2004/07/SoikaAPI.Interfaces.ImgSize">

<a:_x003C_Height_x003E_k__BackingField>высота</a:_x003C_Height_x003E_k__Ba
ckingField>

<a:_x003C_Width_x003E_k__BackingField>ширина</a:_x003C_Width_x003E_k__Ba
ckingField>
        </Size>
        <Type>тип изображения: .png, .jpg, .tiff</Type>
</Img>
<Name>наименование профиля</Name>
<PIndex>индекс страницы</PIndex>
<PostSkewAngle>угол устранения перекоса</PostSkewAngle>
<RecognizeProfileId>идентификатор профиля распознавания</RecognizeProfileId>
<RepresentationParams xmlns:a="http://schemas.datacontract.org/2004/07/SOICAII"/>
<RepresentationType>тип репрезентации: MAIN (основная)</RepresentationType>
<TextLines>
        <ItemList>
                <TextLine>
                        <Height>высота</Height>
                        <Width>ширина</Width>
                        <X>x-координата</X>
                        <Y>y-координата</Y>
                        <Confidence>конфиденс текстовой линии</Confidence>
```

```
                    <Text>Текст</Text>
                    <BlockType>Тип текстового блока: CaptionText</BlockType>
                    <Id>идентификатор линии</Id>
                    <IndexPage>индекс линии на странице</IndexPage>
                    <Words><!--слова-->
                        <ItemList>
                            <Word>
                                <Height>высота</Height>
                                <Width>ширина</Width>
                                <X>x-координата</X>
                                <Y>y-координата</Y>
                                <Confidence>конфиденс
слова</Confidence>
                                <Text>Текст</Text>
                                <Angle>угол</Angle>
                                <Id>идентификатор</Id>
                                <IndexLine>индекс слова в
линии</IndexLine>
                                <IndexPage>индекс слова на
странице</IndexPage>
                                <IsAbbreviation>false</IsAbbreviation>
                                <IsCity>false</IsCity>

        <IsCompanyName>false</IsCompanyName>
                                <IsMail>false</IsMail>
                                <IsName>false</IsName>
                                <IsStreet>false</IsStreet>
                                <KnownColor>Цвет:
GRAY</KnownColor>
                                <Language>Язык: rus</Language>
                                <RealH>реальная высота символов слова в
долях от высоты страницы</RealH>
                                <RealHeight>реальная высота символов
слова</RealHeight>
                                <Symbols>
                                    <ItemList/>
                                </Symbols>
                                <WordType>тип слова:
LETTERS</WordType>
                            </Word>
                        </ItemList>
                    </Words>
                </TextLine>
            </ItemList>
        </TextLines>
</RecognizeResult>
```

Тип блока PolyBlockType:

```
    Unknown = 0,
    FlowingText = 1,
    HeadingText = 2,
    PullOutText = 3,
```

Equation = 4,
InlineEquation = 5,
Table = 6,
VerticalText = 7,
CaptionText = 8,
FlowingImage = 9,
HeadingImage = 10,
PullOutImage = 11,
HorizontalLine = 12,
VerticalLine = 13,
Noise = 14,
Count = 15

Тип репрезентации RepresentationTypes:

MAIN = 0,
BLOCKS = 1

Rest-сервис управления пакетами

Базовый адрес: http://{адрес сервера}/SoicaService/BatchStatusViewer.svc

Сервис управления пакетами позволяет отслеживать статусы пакетов, удалять пакеты, отправлять пакеты на доработку, получать обобщенную информацию о пакете и владельце пакета.

1. Получить статус пакета: /BatchState/{batchGuid}, где-batchGuid - guid пакета.
   Http-метод: GET
   Возвращает строку, содержащую название модуля, в котором в данный момент находится пакет:
   import – импорт
   recognize – распознавание
   validation – валидация
   export – экспорт
   deleted - пакет был удален
   inaccessible - пакет недоступен
2. Получить описание пакета: /BatchDescription/{batchGuid}, где-batchGuid - guid пакета.
   Http-метод: GET
   Результат:
   {"ClassName":"string", // наименование класса пакета
   "CreationDate":"string", // дата в формате dd.mm.yyyy
   "CreationTime":"string", // время в формате hh:mm
   "IsLocked":bool, // признак блокировки пакета true или false
   "Module":"string", // отображаемое наименование модуля
   "State":"string", // отображаемое состояние пакета
   }
3. Получить информацию и статусы всех доступных пакетов: /AllBatchesInfo
   Http-метод: GET
   Результат:
   [{"m_Item1": BatchInfo, "m_Item2": статус}]

BatchInfo:
{"ClassName":"string", // класс пакета
"ClassVersion":Int32, // версия класса пакета
"CreationDate":"DateTime", // дата и время пример \/Date(1571923528010+0300)\/
"ErrorMessage":"string", // сообщение об ошибке
"Guid":"string", // guid пакета
"Id":Int32, // идентификатор пакета
"Invalid":Boolean, // true или false
"Name":"string", // наименование пакета
}

4. Получить статусы всех доступных пакетов:
   /AllBatchStates
   Http-метод: GET
   Результат:
   [{"m_Item1":"string", // имя пакета
   "m_Item2":"string", // guid пакета
   "m_Item3":"string" // статус
   }]

5. Получить информацию о пакетах на экспорте и размер сформированных на экспорте файлов:
   /AllBatchExportFileSize/{batchClass}/{exportName},
   {batchClass} – наименование класса пакета,
   {exportName} – название задания экспорта
   Http-метод: GET
   Результат:
   [{"batch_guid":"string", // guid пакета
   "file_size":"string", // размер файла в байтах
   }]

6. Получить guid-ы всех доступных пакетов: /BatchGuids
   Http-метод: GET
   Результат: массив guid-ов пакетов

7. Удалить пакет (при попытке удаления заблокированных пакетов возвращается false):
   /Batch/{guid}
   Http-метод: DELETE
   Результат - true или false (успешность операции удаления)

8. Принудительно удалить пакет (даже если пакет заблокирован):
   /ForceDelete/{guid}
   Http-метод: DELETE
   Результат - true или false (успешность операции удаления)

9. Существует ли имя пакета: /IsBatchNameExist/{name}
   Http-метод: GET
   Результат: значение true или false

10. Получить полный статус пакета: /BatchStateInfo/{batchGuid}, где batchGuid - guid пакета.
    Http-метод: GET
    Результат:
    {

“Module”:”string”, // наименование модуля
“Message”:”string”, // сообщение
“State”:”string”, // статус пакета
“Level”:”string”, // уровень сообщения: Info, Warn, Error, Critical
“Date”:”DateTime” //дата и время
}

11. Получить список guid-ов дочерних комплектов по guid родительского пакета: /GetKits/{guid}, где guid - guid пакета.
Http-метод: GET
Результат: список guid в виде строк

12. Получить статистику по пакетам, импортированным в определенный временной интервал: /GetPackages
Тело запроса:
{
“datefrom”:”string”, // дата от
“dateto”:”string”, // дата до
}
Http-метод: POST
Результат:
{
“validation”:Int32, //число пакетов в модуле Валидация,
“export”:Int32, //число пакетов в модуле Экспорт
}

13. Получить guid-ы пакетов, импортированных в определенный временной интервал и находящихся в модуле: /GetPackageNums/{module}, где module - модуль
Тело запроса:
{
“datefrom”:”string”, // дата от
“dateto”:”string”, // дата до
}
Http-метод: POST
Результат:
{
“packages_count”:Int32, // количество пакетов
“packages_num”:[“string”] // массив guid пакетов
}

14. Получить состояние пакета на момент попадания на валидацию: /GetPackageBase/{guid}, где guid - guid пакета
Http-метод: GET
Результат:
{"package_num":”string”,// guid пакета
"docs_count":int, // количество документов
"docs":[{ // информация по каждому документу
"doc_type": "string", // наименование класса документа
"page_count":int, // количество страниц в документе
"pages":[{"pagenum":int, // номер страницы
"fieldscount":int, // количество полей на странице
"fields":[// поля
{"name":"string", // наименование поля
"value":"string", // значение поля

"trust_value":"float" // степень доверия }]}
]}]
}

15. Получить текущее состояние пакета: /GetPackageWork/{guid}, где guid - guid пакета
Http-метод: GET
Результат:
{"package_num":”string”, // guid пакета
"docs_count":int, // количество документов
"docs":[{ // информация по каждому документу
"doc_type": "string ", // наименование класса документа
"page_count":int, // количество страниц в документе
"pages":[{"pagenum":int, // номер страницы
"fieldscount":int, // количество полей на странице
"fields":[// информация по каждому полю в документе
{"name":"string", // наименование поля
"value":"string", // значение поля
"valid":"bool", // валидность
”approved”:”bool” // подтверждено ли поле в Валидации
}]}
]}]}

16. отправить пакет на доработку в модуль Валидация со статусом Rework и сообщением об ошибке: /Rework/{guid}, где guid - guid пакета
Тело запроса: строка сообщения об ошибке
Http-метод: POST
Результат: отметка об успешности - true или false

17. Получить владельца пакета: /BatchOwner/{guid}, где guid – guid пакета
Http-метод: GET
Результат: информация о пользователе
{
“email”:”string”, // email пользователя
“fio”:”string”, // ФИО
“groups”:[“string”] // список групп, в которых состоит пользователь
}

18. Получить значение свойства пакета*:* /BatchProperty/{guid}/{propName}, где guid – guid пакета, propName – имя свойства
Http-метод: GET
Результат: строка (значение свойства)
Примечание: используя этот запрос, можно получить значения свойств пакета, определенных в сценарии обработки пакета. Например, получить результат сравнения docx-договора с шаблоном договора (изменен-не изменен). В этом случае propName = ‘is_changed’

19. Перевести пакет на другой модуль: /MoveBatchToModule/{batchGuid}/{module}, где batchGuid – guid пакета, module – название модуля,
для модуля валидации – validation
для модуля экспорта - export
Http-метод: GET
Результат: признак успешности операции true или false

Сценарии обработки пакетов

Сценарий 1.

Импорт пакета осуществляется посредством Rest-api, затем пакет распознается и отправляется либо на валидацию, либо на экспорт (зависит от настроек очереди модулей класса пакета).

Шаг 1. Создать пакет, используя функцию http://{адрес сервера}/SoicaService/AutoImport.svc/Create/{BatchClass}/{name}, где BatchClass – наименование класса пакета, name – наименование пакета, картинка передается как поток (см. раздел “Rest-сервис импорта”, п. 1) или функцию: *http://{адрес сервера}/SoicaService/AutoImport.svc/Create/{batchClass}/{name}/{login}*, где batchClass - наименование класса пакета, name - наименование пакета, login - логин пользователя или группы пользователей, картинка передается как поток (см. раздел “Rest-сервис импорта”, п. 2). В случае успешного создания пакета возвращается guid пакета, который необходимо использовать в функциях по дальнейшей работе с пакетом.

Шаг 2. Добавить в пакет файлы: *http://{адрес сервера}/SoicaService/AutoImport.svc/AddToBatch/{batchGuid}*, где batchGuid - guid пакета, картинка передается как поток (см. раздел “Rest-сервис импорта”, п. 3). В пакет можно добавить несколько файлов, вызвав эту функцию несколько раз.

Шаг 3. После формирования пакета, запустить его на обработку функцией: *http://{адрес сервера}/SoicaService/AutoImport.svc/StartProcess?parameters={batchGuid}*, где batchGuid – guid сформированного пакета (см. раздел “Rest-сервис импорта”, п. 4).

Шаг 4. Отслеживать статус пакета: *http://{адрес сервера}/SoicaService/BatchStatusViewer.svc/BatchState/{batchGuid}*, где-batchGuid - guid пакета. Возвращает строку, содержащую название модуля, в котором в данный момент находится пакет (см. раздел “Rest-сервис управления пакетами”, п. 1). Если очередь модулей обработки пакета содержит модуль “Валидация”, то после распознавания пакет попадает на валидацию, валидатор просматривает пакет и при необходимости устраняет ошибки. После валидации пакет передается в модуль “Экспорт”. Если очередь модулей не содержит модуль “Валидация”, пакет после распознавания сразу передается в модуль “Экспорт”.

На этом этапе также можно получить значения свойств пакета, заданных в сценарии обработки пакета: *http://{адрес сервера}/SoicaService/BatchStatusViewer.svc/BatchProperty/{guid}/{propName}* (см. раздел “Rest-сервис управления пакетами”, п. 18)

Шаг 5. Экспорт пакета. После того, как пакет был передан в модуль “Экспорт”, содержимое пакета можно получить функциями экспорта (см. раздел “Rest-сервис экспорта”).

Сценарий 2.

В классе пакета настраивается автоимпорт из почты/папки, пакеты создаются автоматически при выполнении условий автоимпорта, далее пакеты распознаются и отправляются либо на валидацию, либо на экспорт (зависит от настроек очереди модулей в классе пакета).

Шаг 1. Автоимпорт из папки или почты, автоматическое создание пакета, распознавание пакета.

Шаг 2. Получить информацию о доступных пакетах при помощи функций:

1. Получить guid-ы доступных пакетов *http://{адрес сервера}/SoicaService/BatchStatusViewer.svc/BatchGuids*.
   Результат: массив guid доступных пакетов. Имея guid пакетов, можно отслеживать их статусы.
2. Получить краткую информацию по каждому доступному пакету, включая статус пакета *http://{адрес сервера}/SoicaService/BatchStatusViewer.svc/AllBatchStates*
   Результат: массив троек*:*
   [{
   “m_Item1”:”string”, // имя пакета
   ”m_Item2”:”string”, // guid пакета
   ”m_Item3”:”string” // статус
   }]
3. Получить более подробную информацию по каждому доступному пакету *http://{адрес сервера}/SoicaService/BatchStatusViewer.svc/AllBatchesInfo*
   Результат: массив пар
   [{
   “m_Item1”: BatchInfo, // информация о пакете
   “m_Item2”: “string” // статус
   }]
   BatchInfo:
   {"ClassName":"string”, // класс пакета
   "ClassVersion":Int32, // версия класса пакета
   "CreationDate":"DateTime”, // Дата и время пример: \/Date(1571923528010+0300)\/
   "ErrorMessage":”string”, // сообщение об ошибке
   "Guid":"string”, // guid пакета
   "Id":Int32, // идентификатор пакета
   "Invalid":Boolean, // невалидный? true или false,
   "Name":"string” // наименование пакета"}.

Более подробно см. раздел “Rest-сервис управления пакетами”.

Шаг 3. Отслеживать статус конкретного пакета: *http://{адрес сервера}/SoicaService/BatchStatusViewer.svc/BatchState/{batchGuid}*, где-batchGuid - guid пакета. Возвращает строку, содержащую название модуля, в котором в данный момент находится пакет (см. раздел “Rest-сервис управления пакетами”, п. 1). Если очередь модулей обработки пакета содержит модуль “Валидация”, то после распознавания пакет попадает на валидацию, валидатор просматривает пакет и при необходимости устраняет ошибки. После валидации пакет передается в модуль “Экспорт”. Если очередь модулей не содержит модуль “Валидация”, пакет после распознавания сразу передается в модуль “Экспорт”.

На этом этапе также можно получить значения свойств пакета, заданных в сценарии обработки пакета: *http://{адрес сервера}/SoicaService/BatchStatusViewer.svc/BatchProperty/{guid}/{propName}* (см. раздел “Rest-сервис управления пакетами”, п. 18)

Шаг 4. Экспорт пакета. После того, как пакет был передан в модуль “Экспорт”, содержимое пакета можно получить функциями экспорта (см. раздел “Rest-сервис экспорта”)

Сценарий 3.

Импорт пакета для сравнения договора в формате docx с шаблоном договора в формате docx. Для импорта через RestAPI следует использовать:

1. Метод импорта *http://{адрес сервера}/SoicaService/AutoImport.svc /CreateTemplate/{batchClass}/{name}* (см п.6 в разделе Сервис импорта), при этом шаблон договора следует передавать в теле запроса.
2. Договор следует передать вторым запросом: *http://{адрес сервера}/SoicaService/AutoImport.svc/AddToBatch/{batchGuid}* (см п.3 в разделе Сервис импорта).
3. Далее при помощи запроса *http://{адрес сервера}/SoicaService/AutoImport.svc /StartProcess/{batchGuid}* (см п.4 в разделе Сервис импорта) запустить пакет на обработку.
4. Отслеживать статус конкретного пакета: *http://{адрес сервера}/SoicaService/BatchStatusViewer.svc/BatchState/{batchGuid}*, где-batchGuid - guid пакета. Возвращает строку, содержащую название модуля, в котором в данный момент находится пакет (см. раздел “Rest-сервис управления пакетами”, п. 1). Если очередь модулей обработки пакета содержит модуль “Валидация”, то после распознавания пакет попадает на валидацию, валидатор просматривает пакет и при необходимости устраняет ошибки. После валидации пакет передается в модуль “Экспорт”. Если очередь модулей не содержит модуль “Валидация”, пакет после распознавания сразу передается в модуль “Экспорт”.
На этом этапе также можно получить значения свойств пакета, заданных в сценарии обработки пакета: *http://{адрес сервера}/SoicaService/BatchStatusViewer.svc/BatchProperty/{guid}/{propName}* (см. раздел “Rest-сервис управления пакетами”, п. 18)
5. После распознавания и экспорта пакета можно получить список пакетов и размеры файлов на экспорте: *http://{адрес сервера}/SoicaService/BatchStatusViewer.svc/ AllBatchExportFileSize/{batchClass}/{exportName}* (см п.5 в разделе Сервис управления пакетами).
6. Скачать экспортированный файл*: http://{адрес сервера}/SoicaService/BatchStatusViewer.svc/Image/{batchGuid}/{exportName}* (см раздел Сервис экспорта). После поступления запроса на скачивание файла, пакет блокируется и последующие запросы из п.4 данного сценария не будут включать информацию об этом пакете.
7. Удалить пакет после завершения скачивания экспортированного файла: *http://{адрес сервера}/SoicaService/BatchStatusViewer.svc/ForceDelete/{batchGuid}* (см п.8 в разделе Сервис управления пакетами). При помощи этого запроса можно удалить даже заблокированные пакеты.

Использование нейросетей

Некоторыми инструментами в системе производится поиск данных используя натренированные нейросети. Модели TensorFlow позволяют находить такие данные как паспорт, чек, ценник и т д. Тренировка моделей может быть произведена под каждый конкретный случай. Способ тренировки описан здесь: https://habr.com/ru/company/nixsolutions/blog/422353/

Архитектура системы:

По умолчанию сервис работает по адресу: http://localhost/soicaservice

Модуль администратора: http://localhost/administrator

Модуль валидации: http://localhost/validation

Обращения к БД По TCP: PostgreSQL: 5432 MSSQL: 1433.

Импорт из папки и экспорт в папку настраивается в модуле Администратора.

По умолчанию хранилище файлов расположено в папке C:\inetpub\FileStorage, но расположение хранилища файлов может быть изменено в конфигурации модуля сервиса, администратора и валидации, находящейся:

- C:\inetpub\SoicaWebService\web.config
- C:\inetpub\Administrator\web.config
- C:\inetpub\Validation\web.config

Пример настройки расположения FileStorage в конфигурации:

```
<appSettings>
….
   <add key="file_storage_path" value="C:\inetpub\FileStorage" />
….
</appSettings>
```

В качестве хранилища файлов допускается использовать в том числе сетевые папки.

Полная информация о настройке конфигурации:

web.config, секция

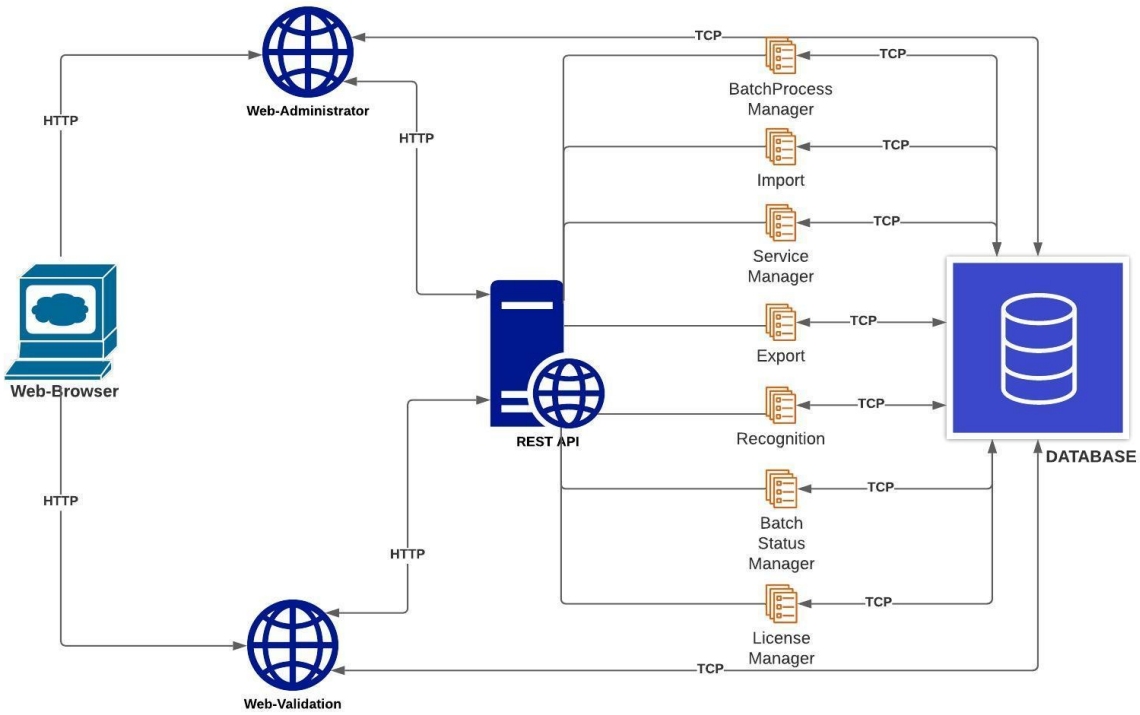
```
<appSettings>
        <add key="Наименование" value="Значение" />
</appSettings>
```

Таблица параметров и значений

№	Наименование	Значение	Значение по умолчанию	Обязательный
1.	DB_path	database=*наименование_БД*;server=127.0.0.1;port=5432;User ID=*пользователь*;password=*пароль*;Timeout=300;CommandTimeout=300;		Да
2.	gs_path_64	Путь с именем файла gsdll64.dll		Да
3.	gs_path_32	Путь с именем файла gsdll32.dll		Нет
4.	lic_path	Путь к файлу с лицензией		Да
5.	server	Адрес master-сервера в распределенной серверной системе. В системе с одним сервером, собственный адрес.	*http://localhost/soicaservice/*	Нет
6.	service_address	Собственный адрес	*http://localhost/soicaservice/*	Нет
7.	DBaseMode	Режим базы данных single (одна) или multiple (много)	*single*	Нет
8.	yaVisionSettingsPath	Только для YandexVision: путь к файлу с настройками.		Нет
9.	soicaII	Путь к библиотеке SOICA_II(CPPDLL).dll		Да
10	soicaII_dict	Путь к файлам словарей	Если не задано, словари при распознавании не используются	Нет
11.	soicaII_dict_size	Максимальный размер загрузки словаря в память	Если не задано, словарь загружается в память целиком	Нет
12.	soicaII_maxCores	Максимальное количество ядер процессора	Рассчитывается в движке SOICAII	Нет
13.	soicaII_coresToBeUsed	Число используемых ядер процессора	Рассчитывается в движке SOICAII	Нет

14.	soicaII_lang	Список загружаемых языков через ‘+’	*rus+eng*	Нет
15.	soicaII_useTorch	Использовать torch (true|false)	*false*	Нет
16.	rec_temp_path	Путь к папке с временными файлами распознавания	*C:\inetpub\ SoicaWebService*	Да
17.	file_storage_path	путь к FileStorage		Да
18.	ram_available	Минимальное количество свободной оперативной памяти во время запуска страниц на распознавание, в Гб, число с плавающей точкой	Если не задано, контроль за памятью не осуществляется	Нет
19.	soicaII_pack_size	Размер пачки при распознавании SOICAII	12	Нет
20.	soicaII_wait_timeout	Интервал ожидания сбора неполной пачки при распознавании SOICAII, в милисекундах	30000	Нет
21.	tess5Path	Путь к библиотеке Tesseract5		Нет
22.	license_server	Адрес master-сервера лицензирования (при использовании распределенной системы лицензирования задается в конфигурационном файле slave-сервера)		Нет
23.	timer_interval	Интервал обхода заданий импорта из папки, в милисекундах	1000	Нет
24.	timer_email_interval	Интервал обхода заданий импорта из почты, в милисекундах	10000	Нет
25.	odata_delay	Задержка между получениями блоков данных с сервера OData	30000	Нет
26.	odata_maxblocksize	Максимальный размер блока при загрузке данных с сервера OData	100000	Нет
27.	odata_timeout	Таймаут ожидания ответа от сервера OData	6000000	Нет

Архитектура SOICA:



Web-Administrator позволяет создавать сценарии, экспортировать и импортировать их для переноса между базами, создавать пользователей и настраивать права доступа для них, настраивать и тестировать распознавание, настраивать источники импорта изображений, осуществлять импорт файлов по сценарию из папки, настраивать пути и структуру экспорта данных, создавать внешние источники и настраивать их.

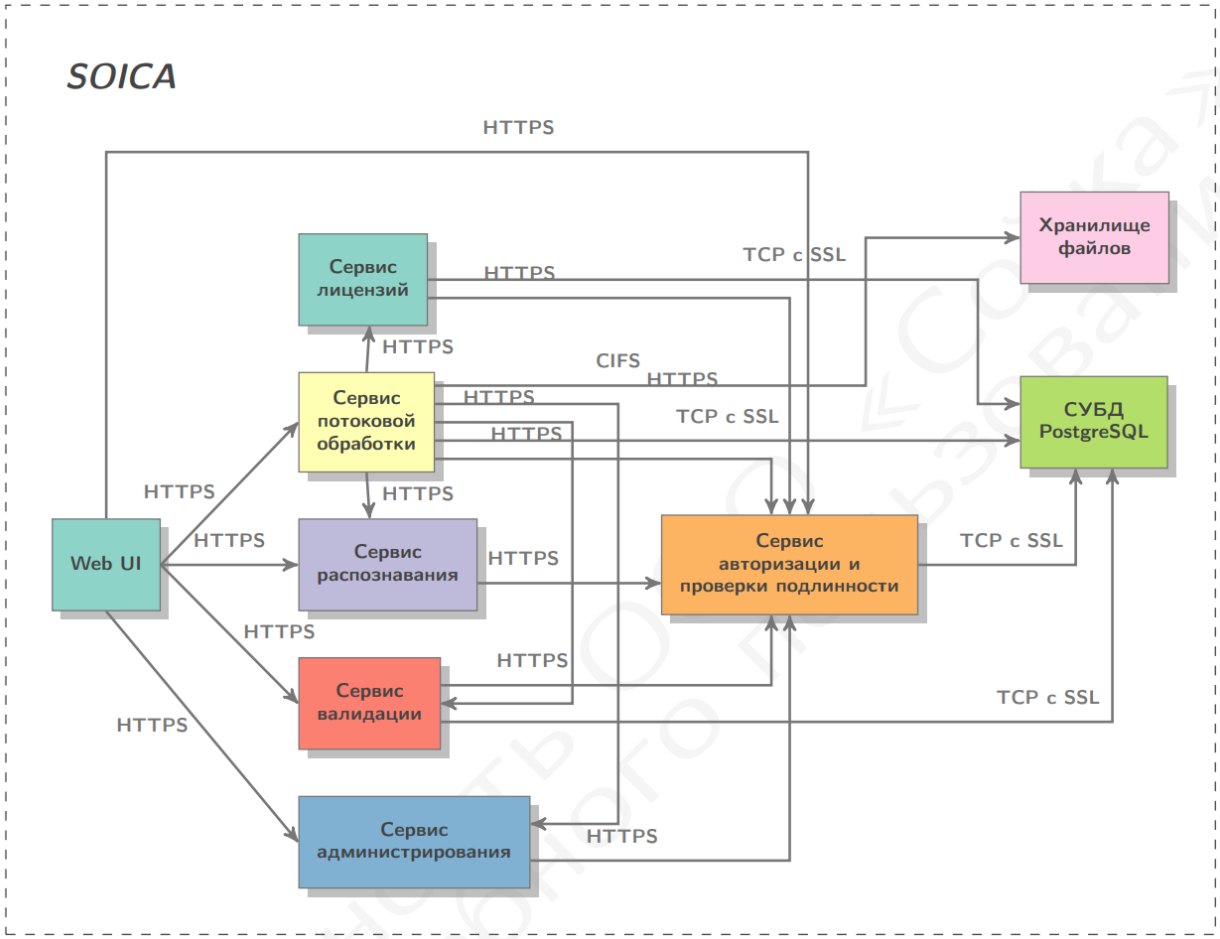
Web-Validation позволяет пользователям проверять распознанные данные и подтверждать их.

REST API сервисы:

- Import – предназначен для загрузки изображений в систему.
- Export – предназначен для получения распознанных данных
- Recognition – предназначен для получения OCR изображения
- License Manager – предназначен для управления лицензиями в случае распределенной установки
- Batch Process Manager – управление процессом обработки пакета
- Service Manager – балансировщик нагрузки с функцией выбора сервиса для обработки пакета
- Batch Status Manager – ручное управление процессом обработки пакета, информирование о состоянии пакета.

Database PostgreSQL или MSSQL. Сценарии могут перемещаться между БД средствами модуля Администратор.

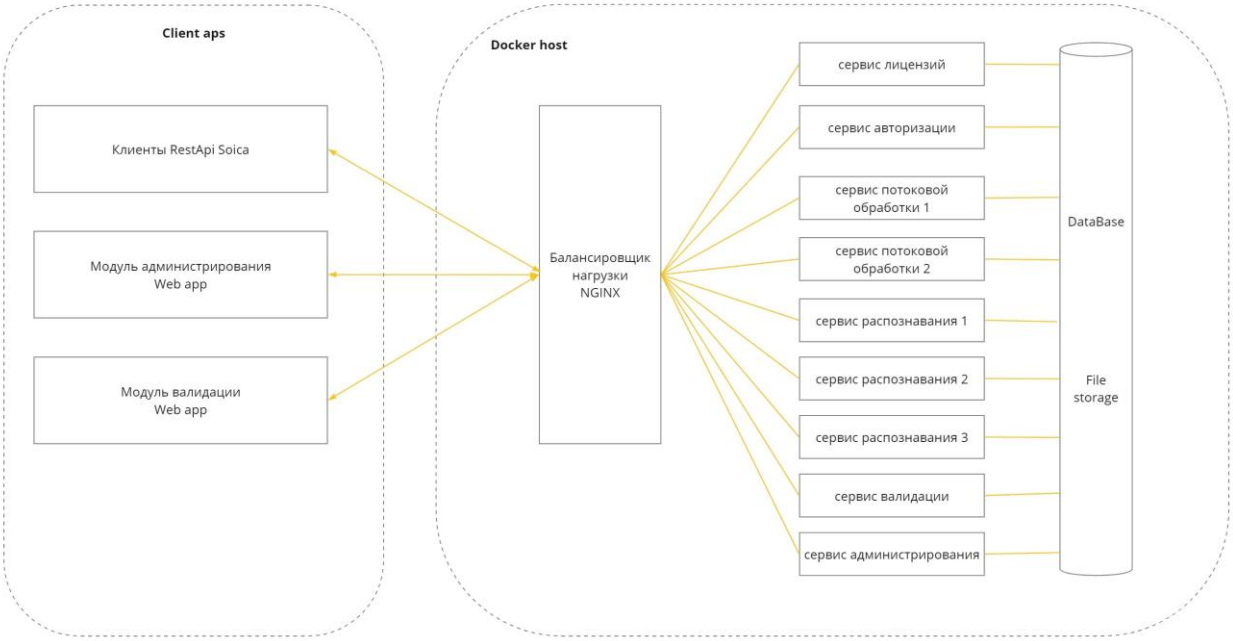
Схема взаимодействия сервисов:



Архитектура распределенной установки:

Каждый сервис SOICA разворачивается в отдельном Docker-контейнере.

Проблема горизонтальной масштабируемости решается добавлением новых экземпляров сервисов в Docker-контейнерах. Балансировка нагрузки выполняется при помощи серверов NGINX.



2. Модуль Администратора.

Модуль предназначен для настройки рабочего проекта.

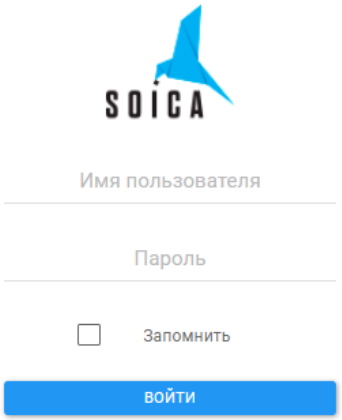
Проект – сценарий от импорта до конечного результата, все профили очистки и распознавания, методы и способы нахождения данных. Каждый проект содержит в себе все этапы обработки. Количество проектов не ограничено. Сам сценарий — это поэтапное выполнение всех модулей приложения.

Внутри проекта следует разделять понятия **Класс пакета** и **Класс документа**.

Класс пакета -это проект в настроечной среде.

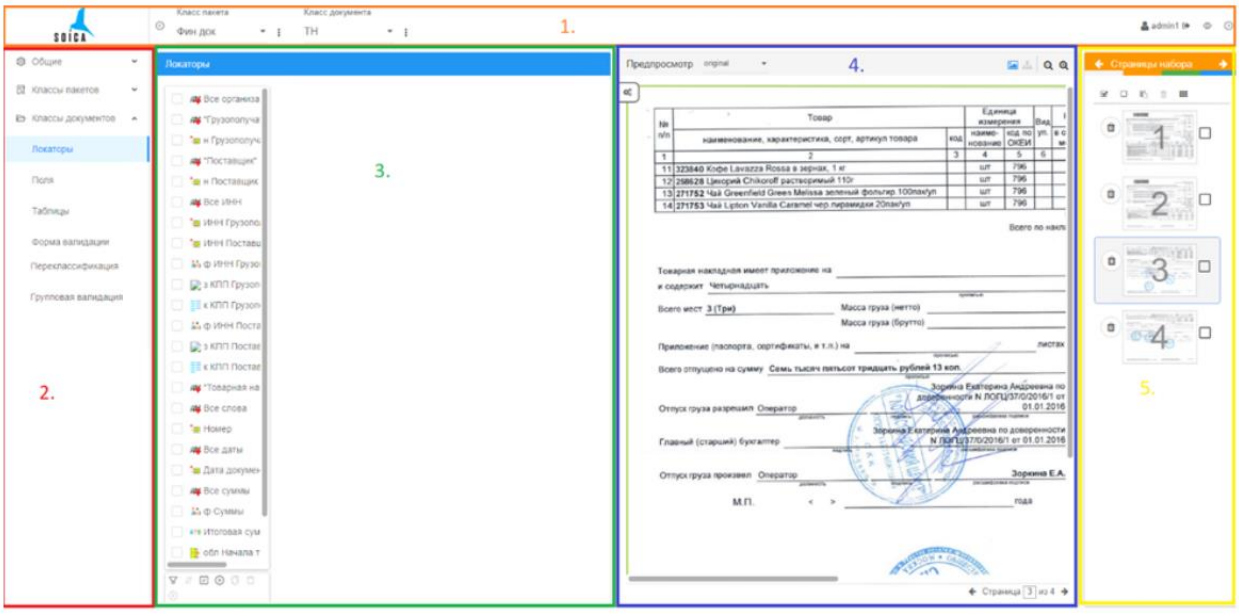
Класс документа относится к конкретному классу пакета и необходим для определения типа изображения. Количество классов документов в проекте **не ограничено.**

Для входа в главное меню модуля пользователю необходимо ввести свой уникальный логин и пароль в соответствующие окна. Далее необходимо нажать кнопку «Войти».



2.1. Интерфейс.

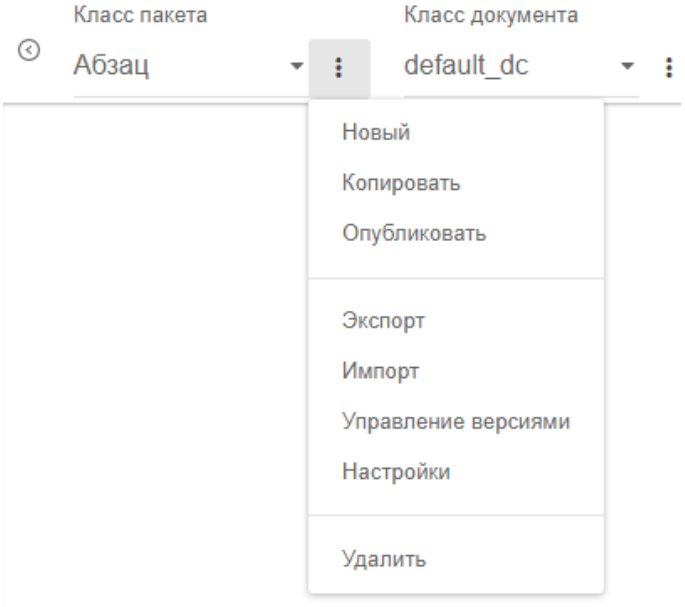
Интерфейс модуля администратора состоит из 5 зон (см. рис 1.). Три из пяти зон могут быть свернуты в зависимости от выполняемой задачи.



(Рис. 1. Общий вид интерфейса)

Зона №1. Панель общей настройки проекта. Панель делится на:

- Логотип «Сойка»;
- Кнопка скрывает зону №2.
- Выбор Класса пакета и Класса документа. В зависимости от выбора класса пакета доступны соответствующие классы документа. Доступные классы пакета добавляются в настройке пользователя.
  Выбор пакета осуществляется из выпадающего списка
  Для создания пакета необходимо открыть меню класса пакета:



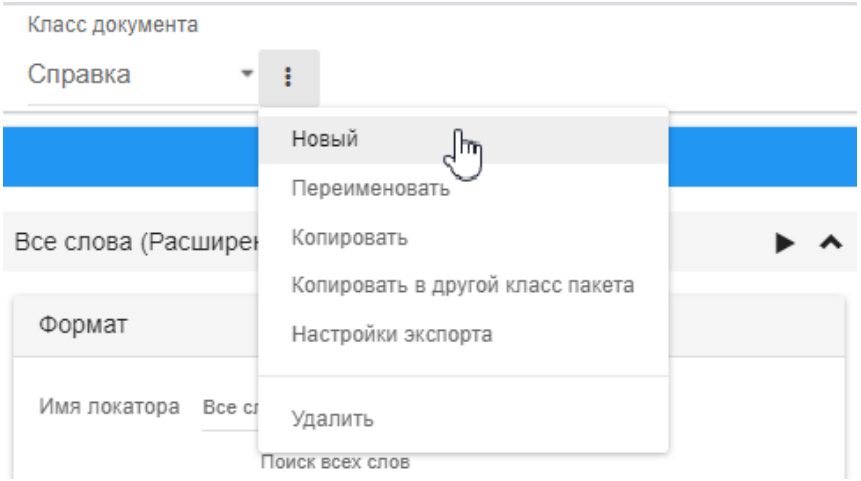
(Рис. 2. Меню Класс пакета)

Описание пунктов меню:

○ **Новый**. При создании нового класса пакета список класса документов будет содержать один класс документа default_dc. В классе пакета не может быть меньше одного класса документа. При создании класса пакета же, создается один профиль распознавания. В классе пакета не может быть менее одного профиля распознавания.

- **Копировать**. Копировать выбранный класс пакета с сохранением всех настроек;
- **Опубликовать**. Для того чтобы внесенные изменения отработали при запуске сценария на сервере, необходимо опубликовать проект.
- **Экспорт**. При запуске экспорта будет создан файл xml. Этот файл можно импортировать в систему на другом сервере.
- **Импорт**. При имеющемся файле класса проекта (xml) можно импортировать заготовленный класс проекта со всеми настройками.
- **Управление версиями**. В рамках выбранного класса пакета ведется запись истории публикации класса пакета. С помощью этого пункта меню можно менять опубликованные версии пакета.
- **Удалить**. Удалить класс пакета из системы.

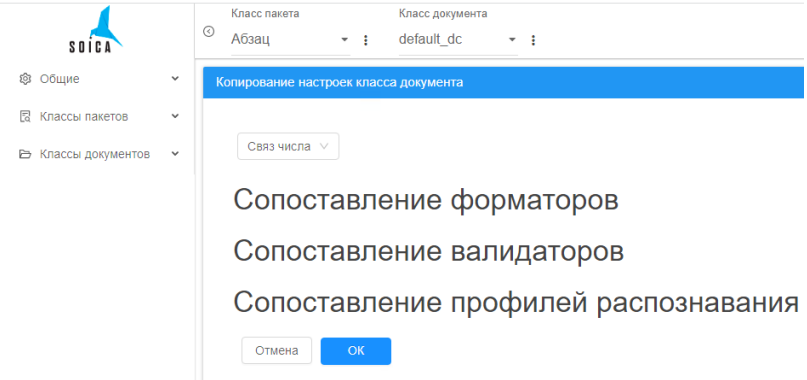
Выбор класса документа осуществляется с помощью кнопки
Для создания класса документа необходимо открыть меню:



(Рис. 3. Меню Класс документа)

В меню можно:

- Новый. Создание нового класса документа. Класс пакета не изменяется.
- Переименовать. Каждый класс документа можно переименовать на любом этапе работы над проектом.
- Копировать. Класс документа можно копировать с сохранением всех настроек. Копия класса документа будет внутри выбранного класса пакета.
- Копировать в другой класс пакета. Класс документа можно копировать с сохранением всех настроек в другой класс пакета. При этом нужно произвести сопоставление форматоров, форматоров, валидаторов и профилей распознавания.



- Настройки экспорта. Для каждого класса документа можно настроить отдельный сценарий экспорта в зависимости от ТЗ заказчика. Каждый тип документов может экспортироваться в разном виде, а также один класс документа может экспортироваться несколькими способами.
- Удалить. Класс документа будет удален из класса пакета с удалением всех настроек.

- Имя пользователя, под которым выполнен вход.
- Кнопка скрывает зону №4.
- Кнопка скрывает зону №5.

Зона №2. Навигационное меню. Служит для изменения рабочей области и выбора этапа сценария для внесения изменений в проект.

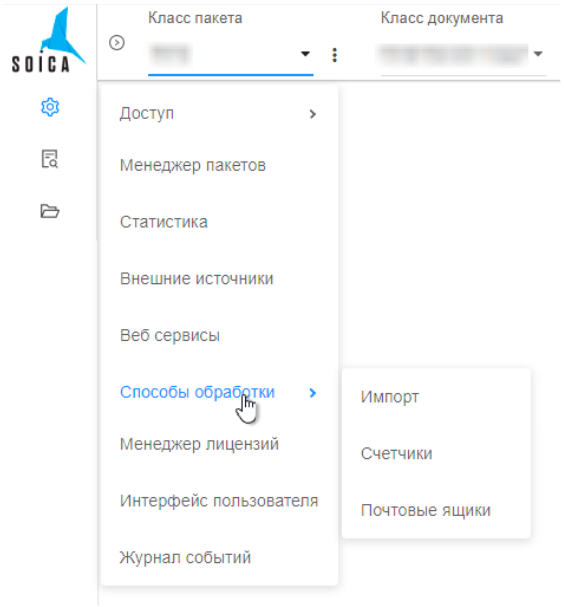
Зона №3. Рабочая область настроек. Служит для работы над настройками системы.

Зона №4. Меню просмотра результатов. Служит для отображения выбранного изображения и наглядного представления внесенных изменений в проект.

Зона №5. Меню набора изображений. Служит для добавления файла в доксет; быстрого просмотра изображения и выбора конкретного изображения для применения внесенных изменений в проекте; отображения результата работы классификации.

2.1.1. Навигационное меню.

Навигационное меню служит для изменения рабочей области и выбора этапа сценария для внесения изменений в проект. Меню делится на 3 части. При раскрытии одной части другая сворачивается. В свернутом состоянии каждая часть меню отображается в виде всплывающей формы справа от иконки части меню при наведении на него курсора.



2.1.2. Рабочая область настроек.

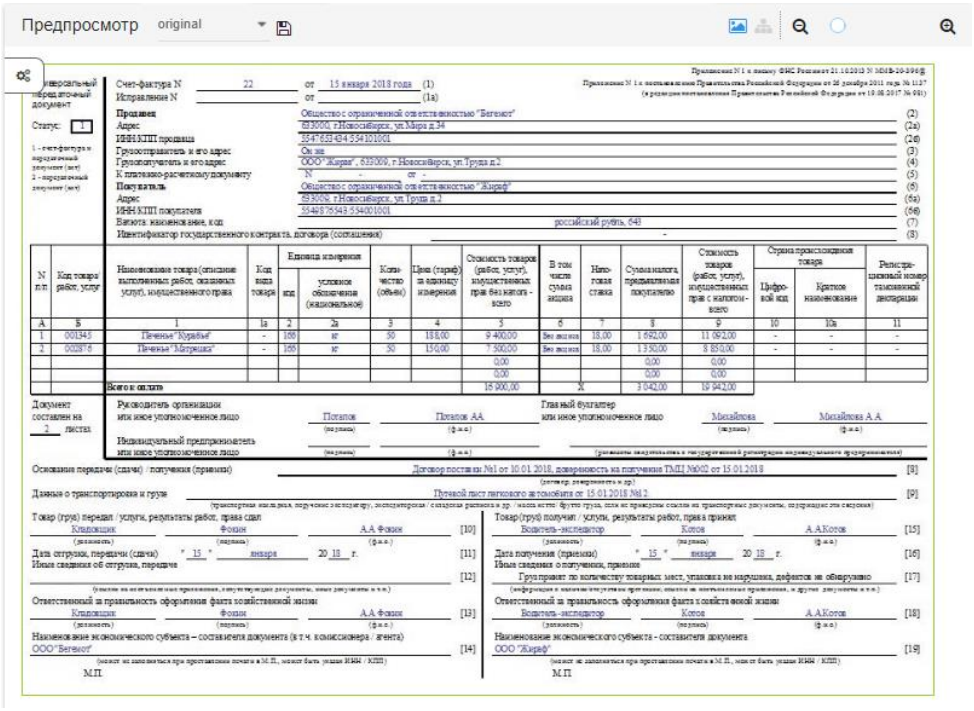
Рабочая область настроек служит для отображения инструментов, выбранных в навигационном меню. В левом верхнем углу можно увидеть название элемента, с которым ведется работа. Рабочую область настроек нельзя скрыть.

2.1.3. Меню просмотра результатов.

Меню просмотра результатов меняется в зависимости от выбранного изображения.

Существует два вида просмотра результатов:

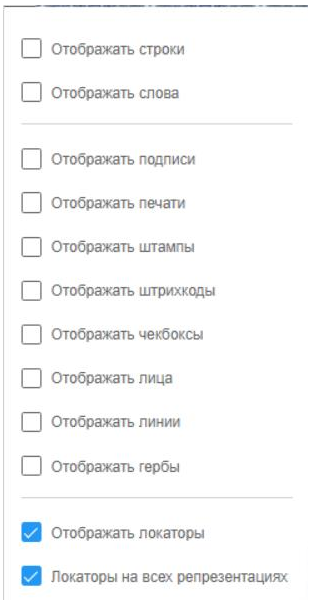
- **Графический предпросмотр. ( )**



(Рис. 4. Меню просмотра изображения (графическое отображение))

При выборе данного вида просмотра изображения можно увидеть какие профили распознавания были применены к документу и как они выглядят;

У этого вида предпросмотра есть свое меню:



(Рис. 6. Меню графического предпросмотра)

Отображать строки – на изображении будут выделены строки распознавания текста в зависимости от выбранного профиля распознавания.

Отображать слова – На изображении синим цветом будут выделены результаты OCR в зависимости от выбранного профиля распознавания.

Опции меню «Отображать подписи», «Отображать печати», «Отображать штампы» и «Отображать штрих коды» отображают соответствующие элементы, в случае, если они найдены.

Отображать локаторы – в зависимости от выбранного локатора, на изображении синим цветом будет выделен результат его работы.

Локаторы на всех репрезентациях – выбранный локатор будет отображаться на всех профилях распознавания (репрезентациях). Если в выбранной репрезентации локатор не нашелся, а на какой-то другой есть результат, то он будет выделен серым.

Визуальное отображение этого меню может меняться с выходом новых версий программы, также количество этих опций может быть больше.

Изображение можно уменьшить или увеличить .

- **Древовидный предпросмотр. ( )**

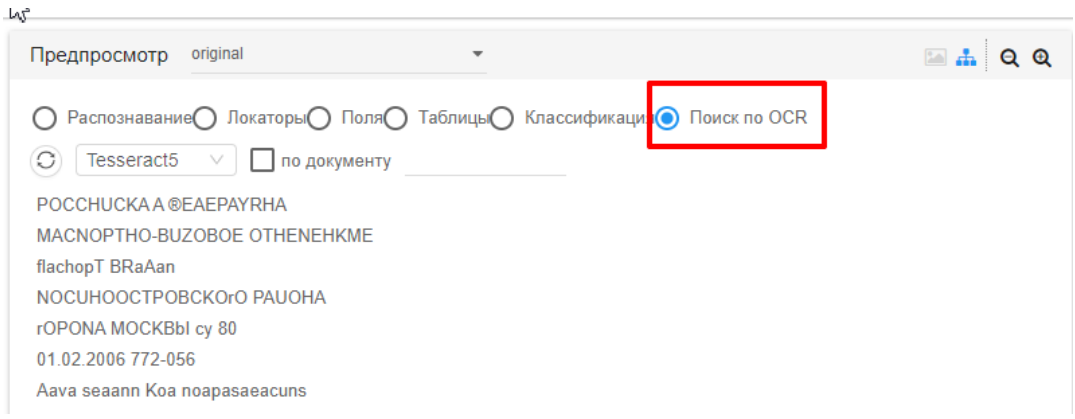
(Рис. 7. Меню просмотра изображения (древовидное отображение))

При древовидном отображении можно увидеть результаты Распознавания (OCR), Локаторов, Полей и таблиц, а также выполненные условия Классификации у выбранного изображения.

При выделении конечного результата в дереве в нижней части можно увидеть координаты, размеры области в которой был найден результат, процент доверия найденного результата, страницу и порядковый номер, профиль распознавания и графическое отображение области.

При нажатии на кнопку в левом нижнем углу можно увидеть найденный результат на всем изображении. Изображение появится на рабочей области настроек.

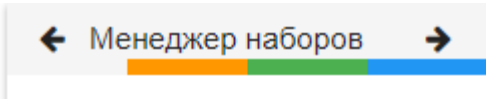
При выборе «Поиск по OCR» можно увидеть текст выбранного профиля распознавания и осуществить поиск по ключевым словам.



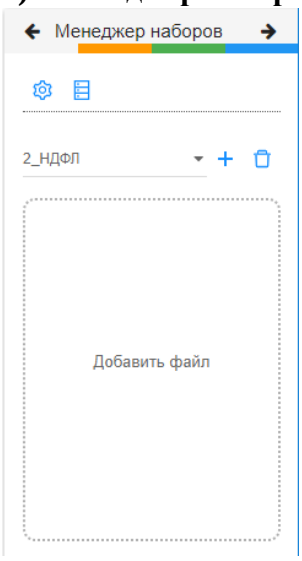
(Рис. 7.1 Поиск по OCR)

2.1.4. Меню набора изображений.

Область работы с набором - инструмент для импорта изображений в набор, их последующей группировки и хранения. Меню делится на 4 вида: Менеджер наборов (серая вкладка), Страницы набора (оранжевая вкладка), Классификация и разделение (зеленая вкладка), Извлечение данных (синяя вкладка).



1) Менеджер наборов



Набор – это группа файлов предназначенных для настройки инструментов поиска данных на примере выбранных изображений.

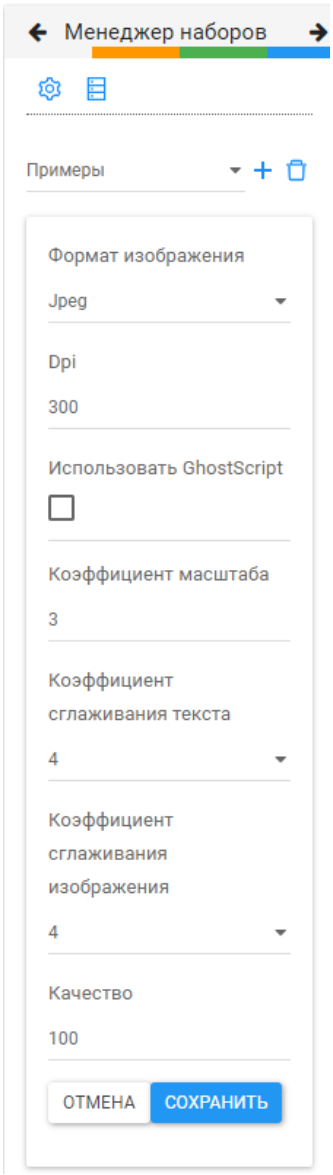
В наборе может быть не ограниченное количество изображений. Набор не зависит от выбранного класса пакета и класса документа, с одним набором можно работать в разных классах пакета.

Для создание нового набора необходимо нажать на кнопку «+», изображения можно добавить через кнопку «Добавить файл», либо «перетащить» файлы в область «Добавить файл».

Поддерживаемые форматы изображений: JPEG, PDF, TIFF, BMP, PNG, DOCX, GIF

При добавлении PDF или TIFF добавляются все страницы из документа.

(Рис. 8. Менеджер наборов)

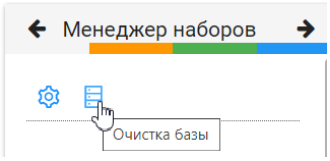


(Рис. 8.1 Настройки для разбора PDF)

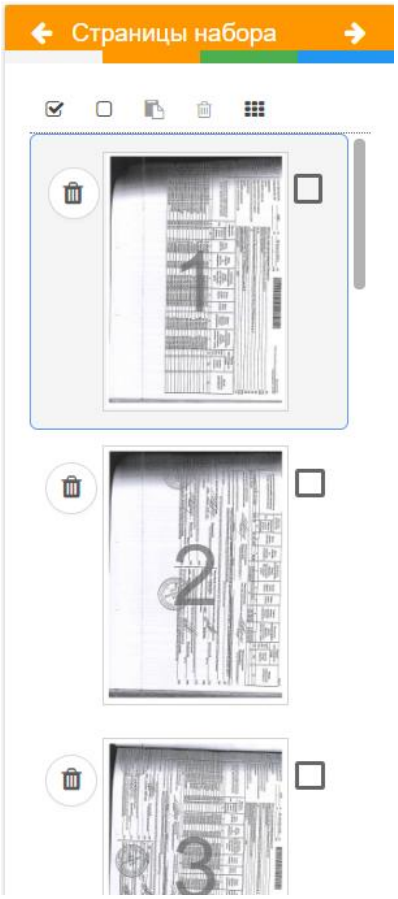
Для страниц в формате PDF можно выбрать в какой формат будут преобразованы страницы поступающего файла (Jpeg, Tiff или PNG). Для этого необходимо нажать на шестеренки в Менеджере наборов. Необходимо указать характеристики обрабатываемого изображения: Dpi преобразованного изображения (диапазон он 72 до 600); Качество jpg (только для jpg, от 0 до 100); Тип сжатия tiff (только для tiff).

Все пользователи видят созданные наборы, а также при удалении набора он пропадет у всех пользователей системы.

Правее «шестерёнки» расположена кнопка «Очистка базы», позволяющая удалять непривязанные к пакетам или доксетам страницы, репрезентации, и прочее.



2) Страницы набора.



После загрузки файлов каждый из них делится на отдельные страницы.

Каждая страница получает свой порядковый номер в наборе.

В виде «Страницы набора» можно отмечать сразу несколько страниц и выполнять над ними один этап обработки (очистка, распознавание)

Из страниц набора можно удалять страницы по одной, либо сразу несколько.

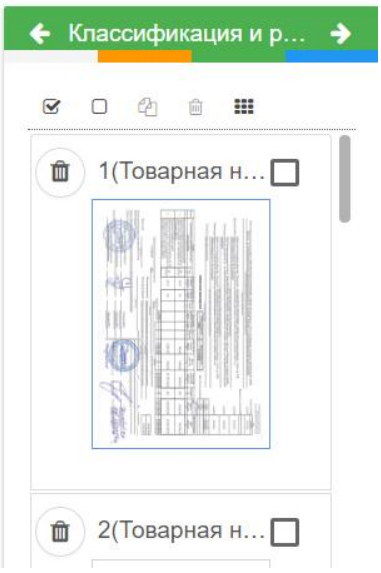
Для того чтобы страница отобразилась в области предпросмотра необходимо выделить саму страницу.

- Выбор вида отображения. Предусмотрено 3 вида отображения страниц в списке: Крупные значки; Мелкие значки; Список.

- Создать документ для извлечения данных. Каждую страницу можно перенести в отдельный документ выбранного класса минуя автоматическую классификацию. Передавать страницы можно по одной, либо несколько.

(Рис. 9. Страницы набора)

3) Классификация и разделение.



Страницы после выполнения автоматической классификации формируются в документы. В одном документе может быть несколько страниц.

В виде классификация и разделение отображаются распознанные и класифицированные документы.

Документы можно копировать в извлечение данных по одному, либо несколько сразу ( ).

Удалять так же можно сразу несколько документов.

Предусмотрено отображение документов 2 способами Крупные значки и Список.

Этот вид используется для настройки автоматической классификации.

(Рис. 10. Классификация и разделение.)

4) Извлечение данных

(Рис. 11. Извлечение данных)

Вид «Извлечение данных» основной для работы и настройки извлечения данных.

В этом виде отображаются верно классифицированные и распознанные документы.

Предусмотрено отображение документов 2 способами Крупные значки и Список.

Удалять документы можно по одному или несколько сразу.

Пока документ находится в списке его нельзя удалить из набора «Страницы набора».

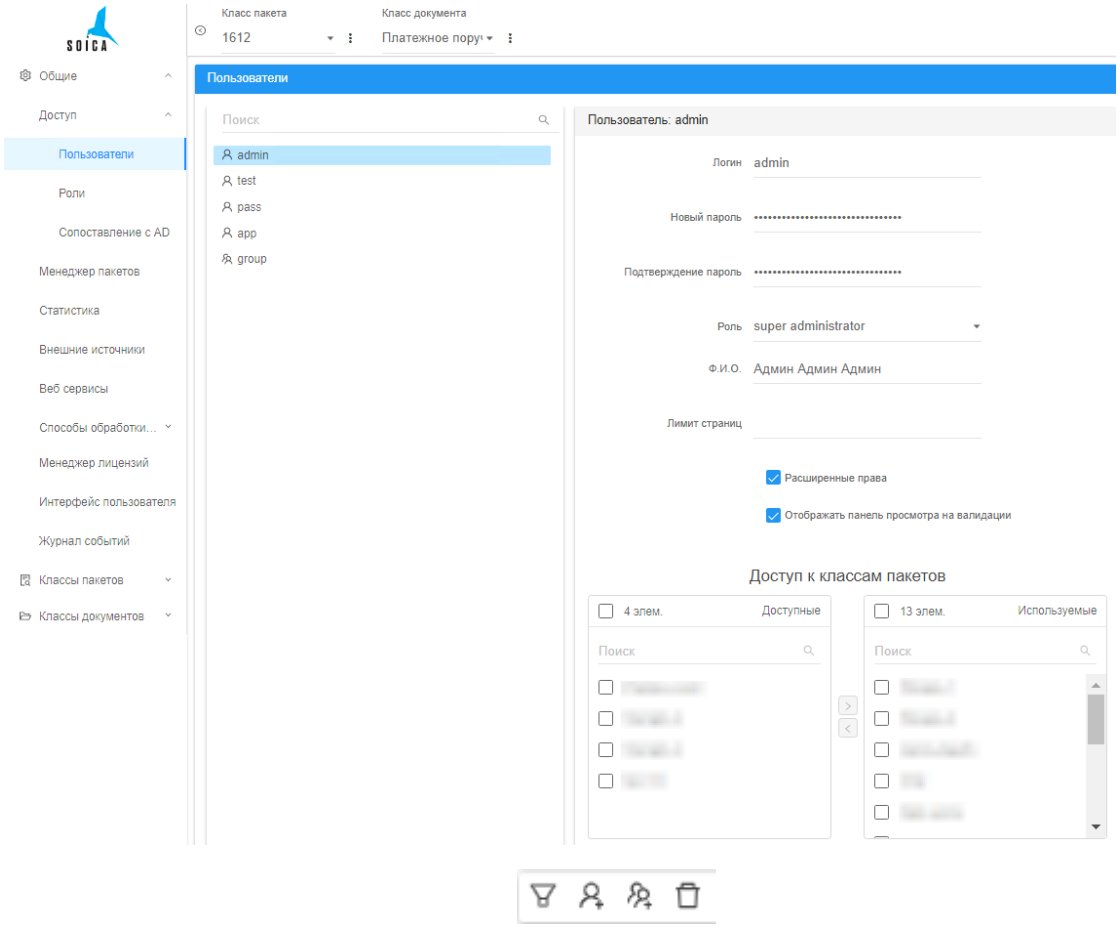
Настройка локаторов, полей, таблиц происходит именно на документах из вкладки «Извлечение данных»

Если документу необходимо переполучить профиль распознавания, то данная операция выполняется на вкладке «Страницы набора». Повторно добавлять перераспознанную страницу в документы не нужно. Репрезентация соответствующего документа на вкладке «Извлечение данных» автоматически обновится.

2.2. Меню настройки (Общие).

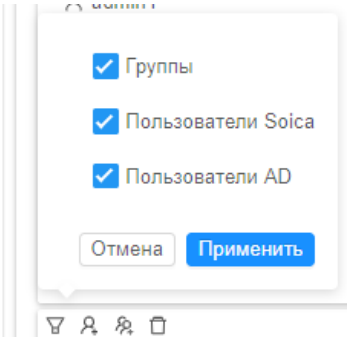
2.2.1. Доступ. Пользователи.

В общем списке пользователей виден логин и права имеющихся пользователей.



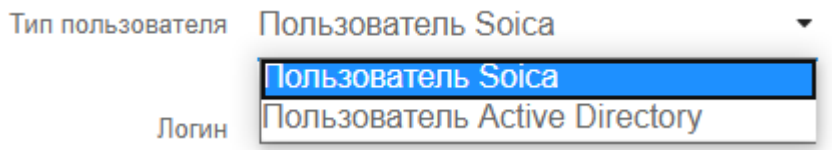
Из области управления списком можно

- С помощью фильтра можно увидеть список пользователей AD или список групп.



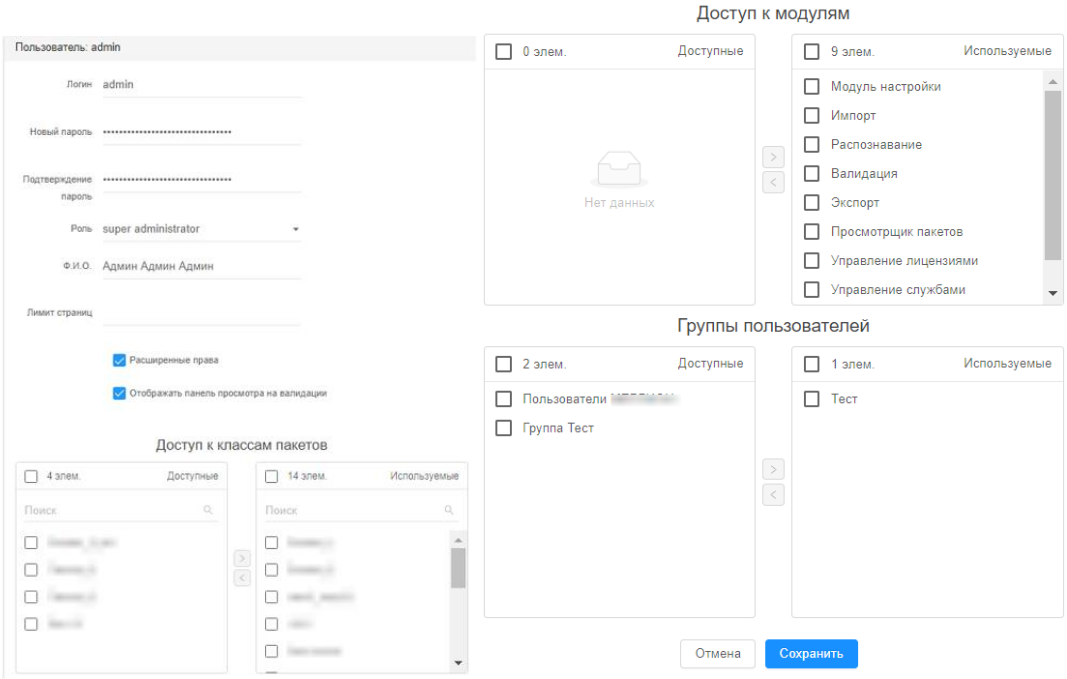
- Создать нового пользователя.
- Добавлять новую группу.
- Удалить выбранного пользователя.

При создании пользователя можно выбрать связь системы с Active Directory.



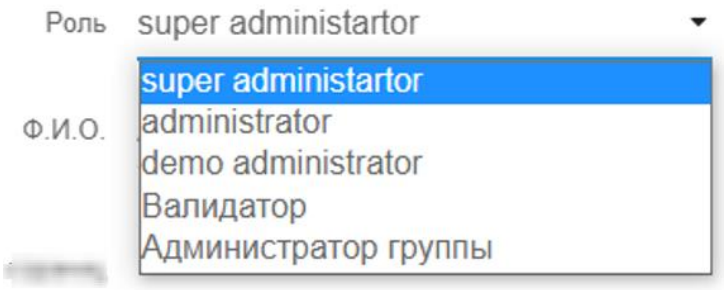
Выбрав пользователя в списке мы можем перейти к его редактированию.

Редактирование пользователя:



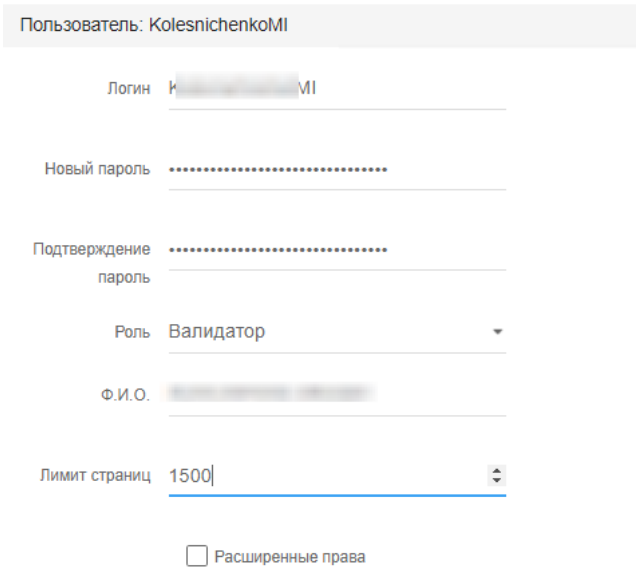
(Рис. 12.1 Редактирование пользователя.)

В области редактирования пользователя можно сменить Логин и пароль пользователя, роль. Базовые роли: super administrator, administrator, demo administrator, Валидатор, Администратор группы. Инженер может создать новую роль с уникальным именем и своим уникальным набором настроек.



Задать лимит загружаемых страниц. Установить «Расширенные права».

Лимит лицензий, можно задать для каждого пользователя в системе, указав количество страниц, который он может обработать.



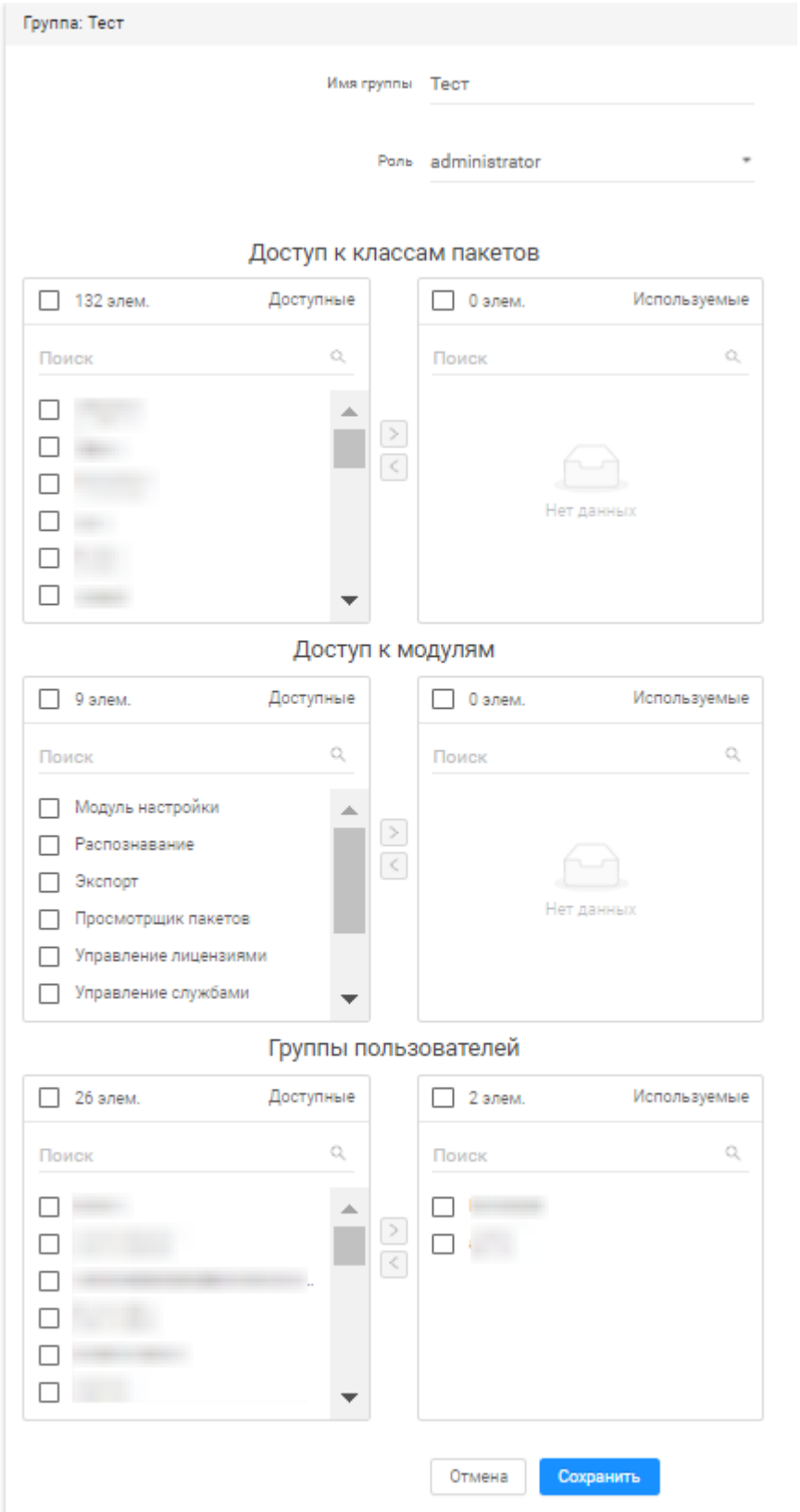
Роли:

- **Super administrator** – полные права на весь модуль администратора.
- **Administrator** – все права на модуль администратора, кроме создание и редактирование данных пользователей и групп пользователей, есть права только на просмотр списка пользователей.
- **Demo administrator** – права только на создание и редактирование объектов из меню «Класс пакета» и «Класс документа». Редактировать может только те объекты, которые сам создал. Имеет доступ к просмотру менеджера пакетов.
- **Валидатор** – права на открытие пакетов внутри модуля валидации, которые назначены на этого пользователя, либо назначены на группу, в которую входит этот пользователь.
- **Администратор группы** – права на открытие пакетов внутри модуля валидации, которые назначены на всех пользователей группы, администратором которой является, а также имеет права на создание и редактирование пользователей группы.

Для добавления пользователю класса пакета необходимо найти пакет в списке «Доступные» и перенести в список «Используемые».

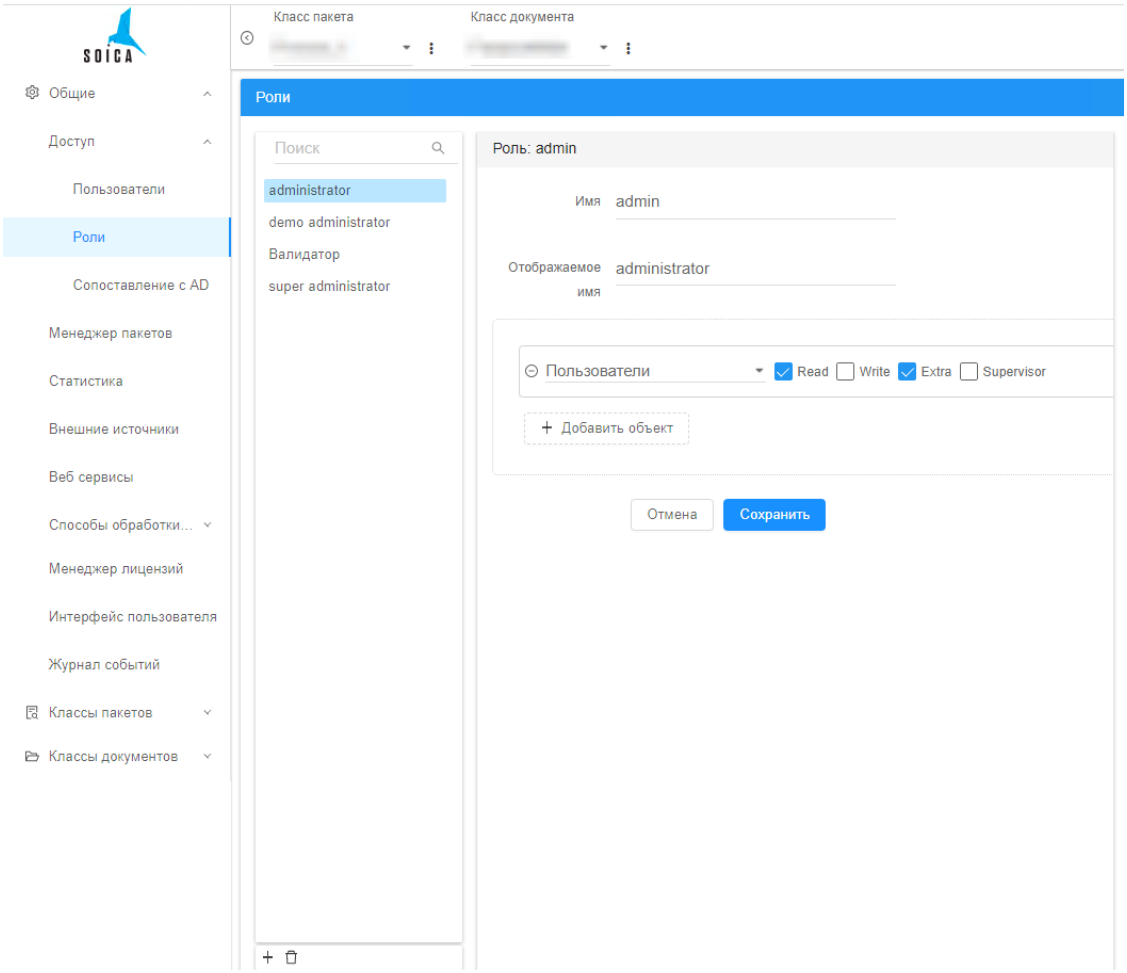
Аналогично происходит управление доступом к модулям системы.

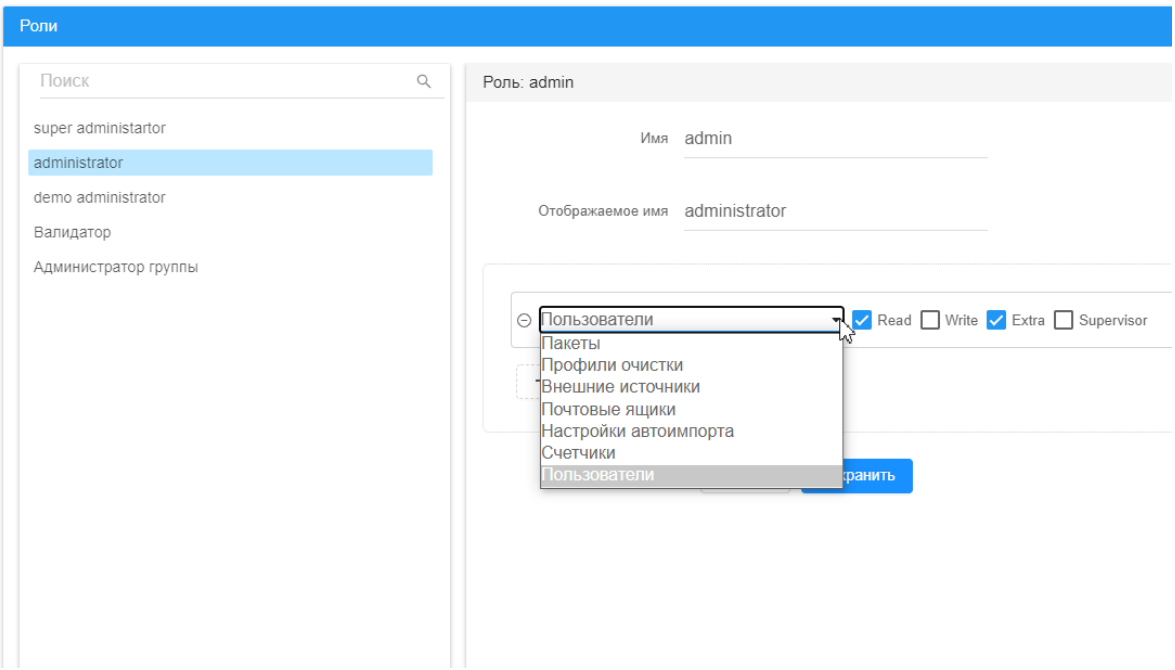
Система позволяет создавать группы пользователей и устанавливать им одинаковые права. Управление группами аналогично классам и модулям. Создание и управление группами происходит методом добавления пользователей из одного списка в другой. В группу может быть добавлен только пользователь. Группу из групп делать нельзя.



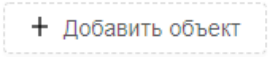
2.2.2. Доступ. Роли.

Управление доступами осуществляется в разделе Роли. Если не выбран ни один объект, доступ будет осуществлен ко всем объектам, в рамках прав выбранной роли.





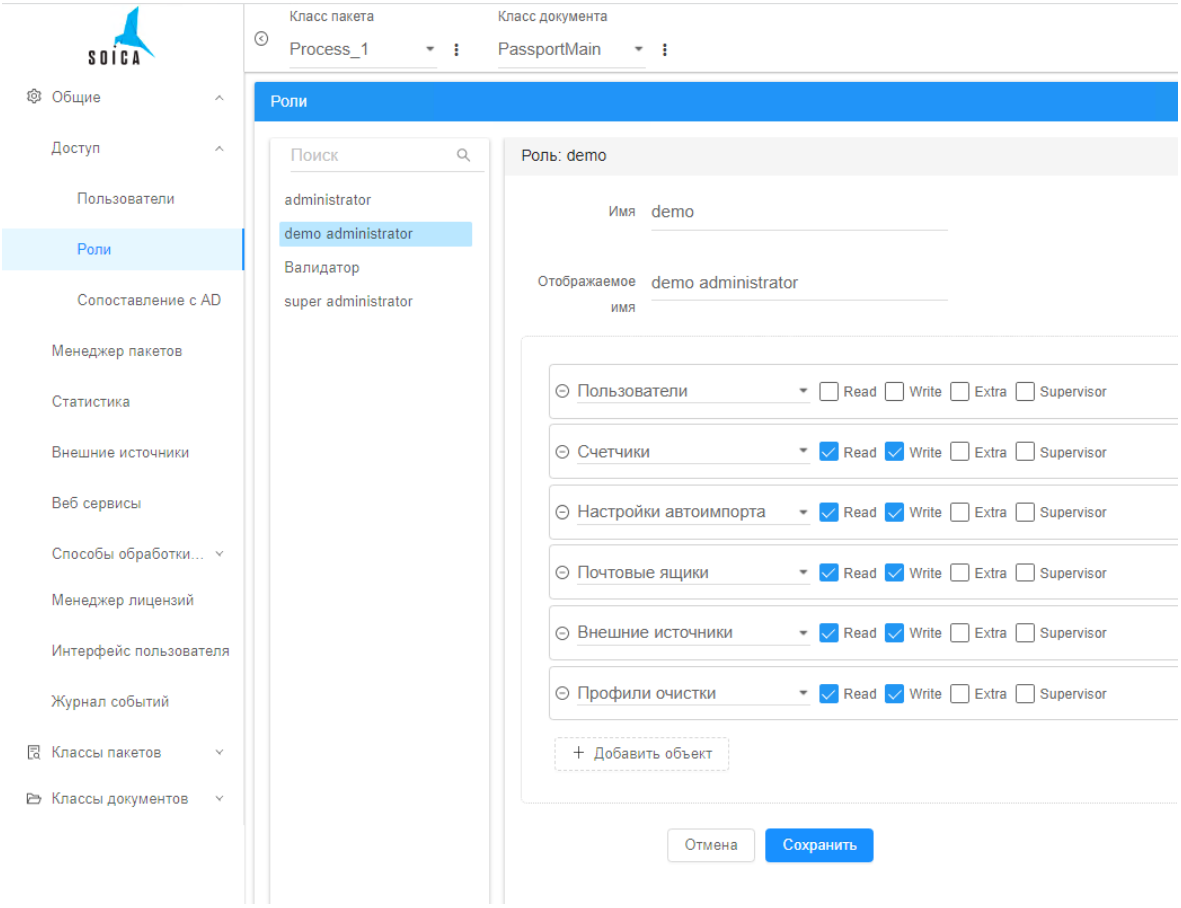
В качестве доступного объекта можно выбрать: Пользователи, Пакеты, Профили очистки, Внешние источники, Почтовые ящики, Настройки автоимпорта, Счетчики,

Пользователи. Добавление объекта для настройки прав доступа к нему производится нажатием кнопки 

.

Возможные действия над выбранным объектом:

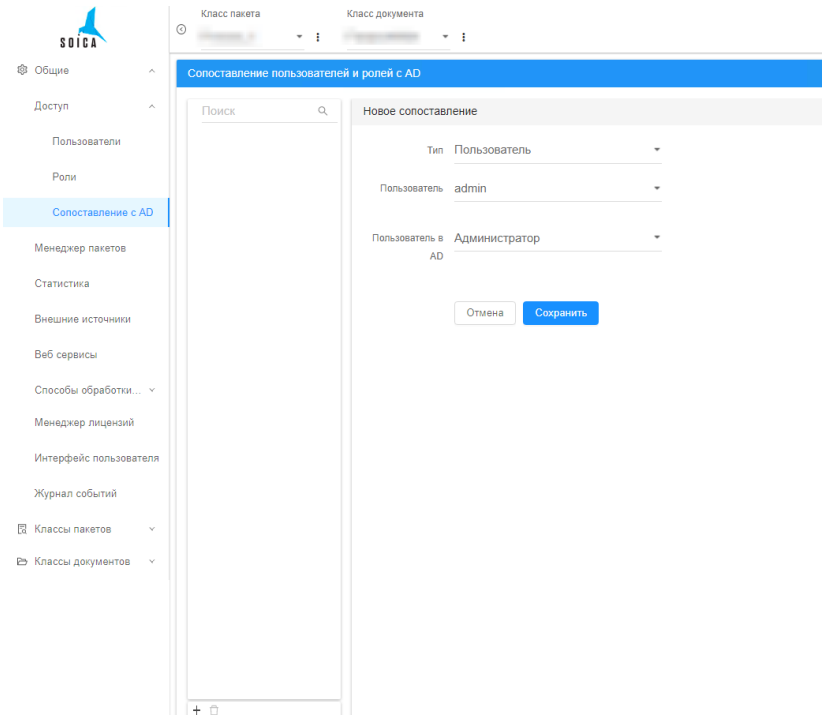
- **Read** – просмотр всех объектов, без права вносить изменения.
- **Write –** может вносить изменения только в те пакеты, которые создал сам.
- **Extra –** может вносить изменения во все пакеты.
- **Supervisor –** если находится в группе пользователей, то видит все объекты группы и может управлять ими в зависимости от выбранных опций.



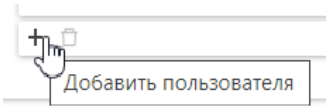
Количество создаваемы пользователей не ограничено.

2.2.3 Доступ. Сопоставление с AD

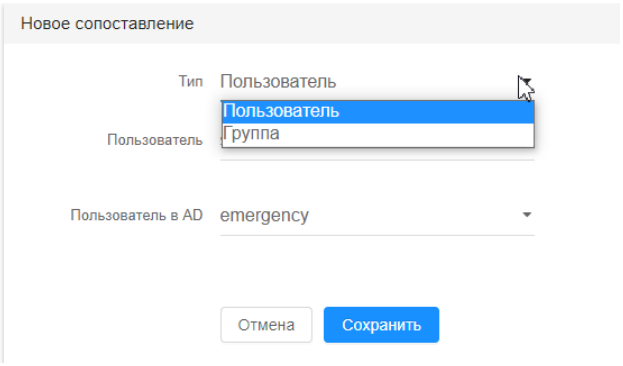
Выполняется сопоставление учетных записей, созданных в Soica и в Active Directory.



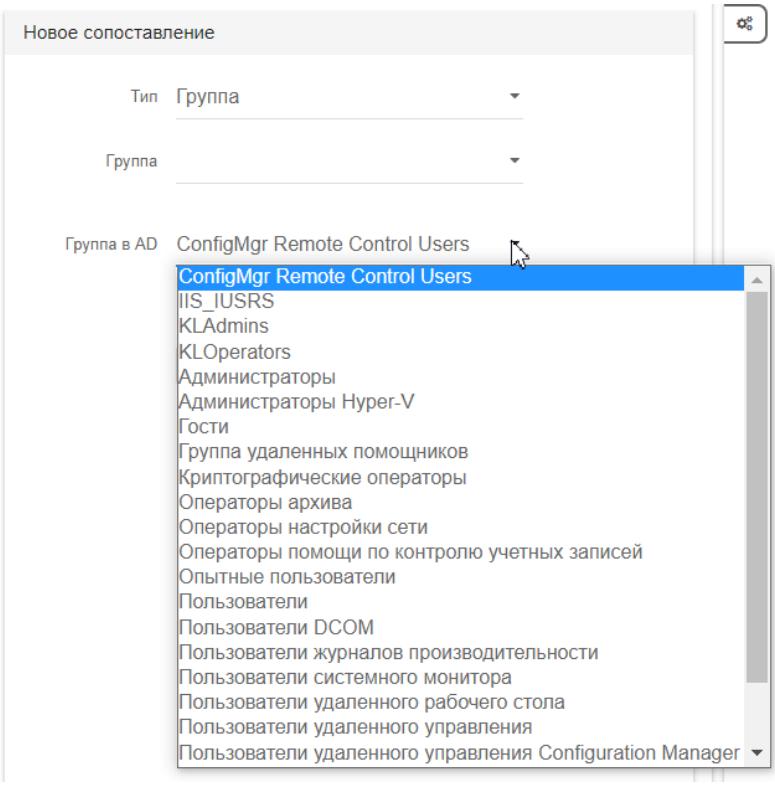
Добавить пользователя можно нажатием «+»



В поле Тип можно выбрать пользователя или группу пользователей.

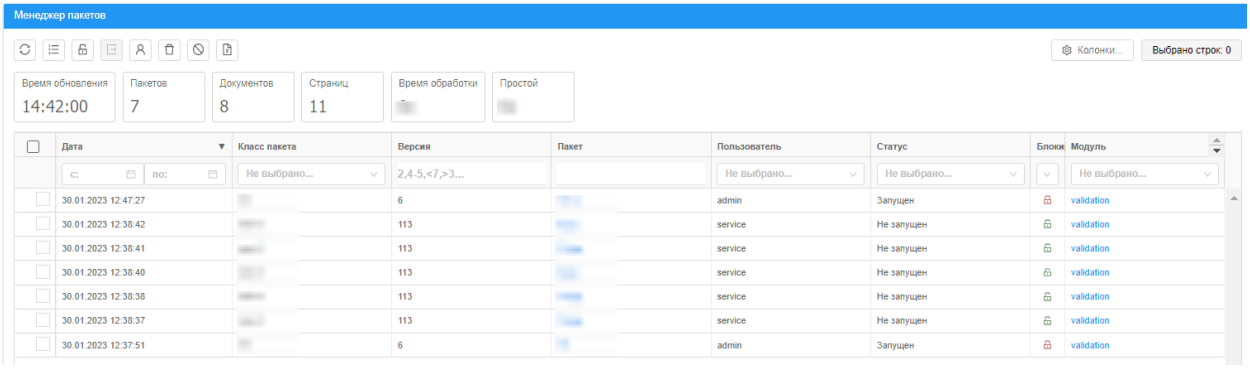


При выборе типа Группа необходимо выбрать из списка ранее созданную группу пользователей в Soica, а также выбрать из списка группу в AD



2.2.4 Менеджер пакетов.

Менеджер пакетов необходим для отслеживания состояния пакетов, отправленных по сценарию проекта. В списке пакетов есть возможность фильтрации по каждому столбцу. Столбец «Дата» предусматривает выборку по установленному временному диапазону.



(Рис. 13. Менеджер пакетов)

 - Обновить список пакетов;

 - Изменить приоритет выбранного в списке пакета;

 - Разблокировать выделенные пакеты;

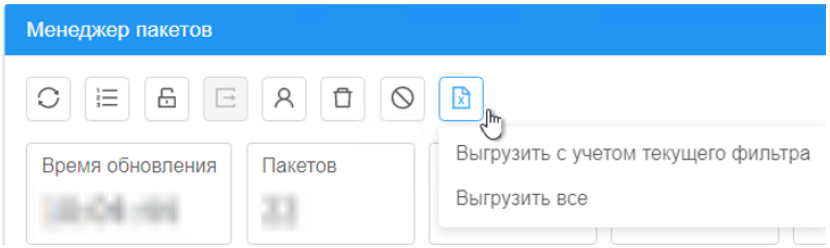
 - Следующий модуль. Переводит выбранный пакет на модуль по выбору;

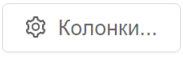
 - Передать пользователю. Позволяет назначить на выделенный пакет конкретного пользователя или группу;

 - Удалить выделенные пакеты;

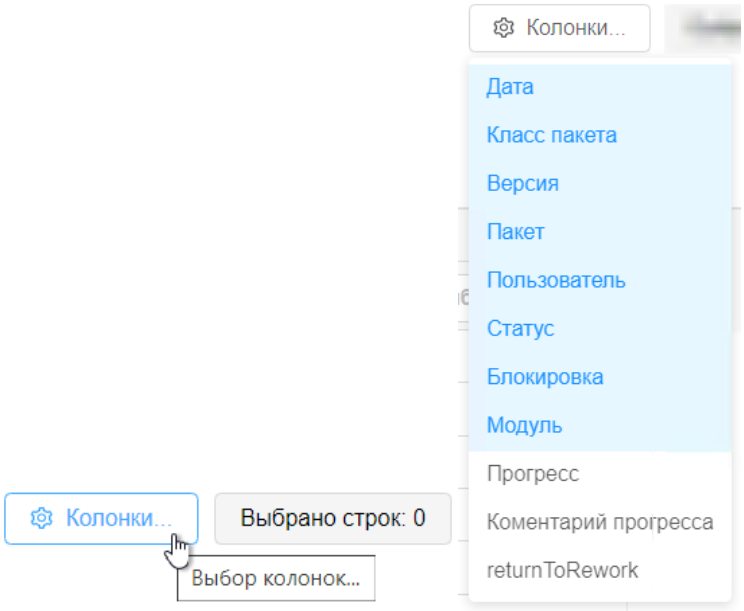
 - Удаляет все пакеты со статусом Завершено, которые располагаются на том модуле, который является последним в очереди модулей для указанного класса пакета.

 - Выгрузить пакеты в формате .XLSX. При нажатии на данную кнопку необходимо выбрать вариант выгрузки. Отфильтрованный ранее список или все пакеты выгружаются в формате .xlsx и автоматически сохраняются. При выборе «Выгрузить с учетом текущего фильтра» выгружаются все пакеты, отфильтрованные в списке пакетов (и выделенные и не выделенные).

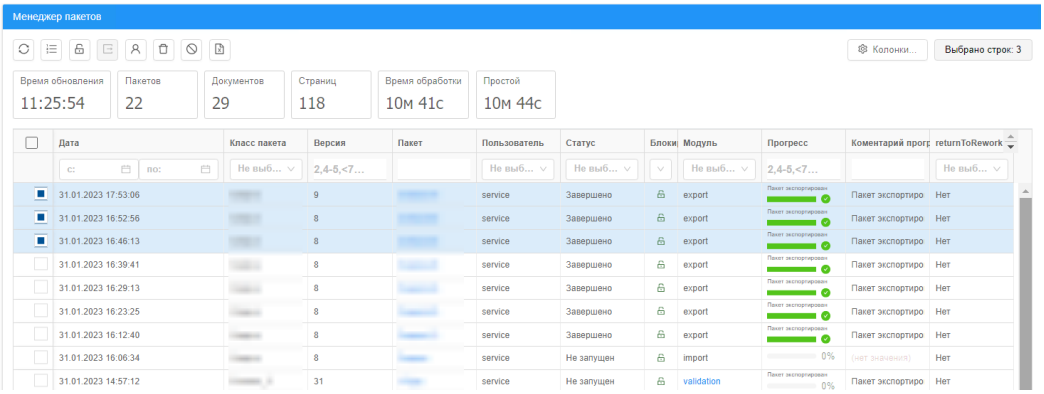




- Выбор отображаемых столбцов, по которым можно производить фильтрацию списка пакетов.



- Указывает сколько строк выбрал пользователь.



Панель статистики.

Время обновления- время последнего обновления списка пакетов.

Пакетов - количество пакетов в отфильтрованном списке.

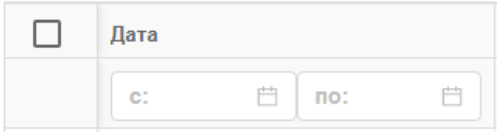
Документов – количество документов в отфильтрованном списке пакетов.

Страниц – количество страниц в пакетах из отфильтрованного списка.

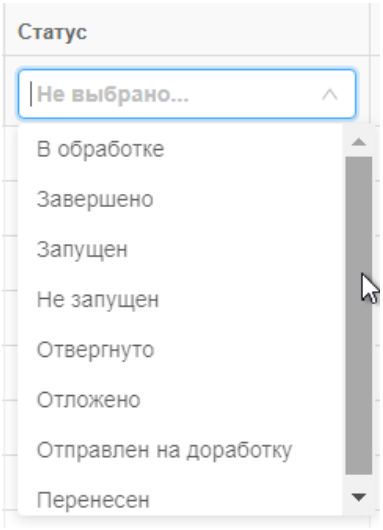
Время обработки – время работы системы начиная с импорта первого пакета в отфильтрованном списке и заканчивая временем передачи последнего пакета в списке на валидацию (либо завершения этапа экспорта).

Простой – время, в течении которого не производилась обработка.
Основным элементом является список пакетов. В таблице мы видим:

- **Дата** – столбец дата позволяет сортировать пакеты за определенный период.



- **Класс пакета** - т.к. менеджер пакетов относится ко всей системе, в нем можно увидеть состояние пакетов по всем классам пакетов.
- **Версия** – номер экземпляра Класса пакета (проекта). Порядковый номер версии увеличивается на единицу после каждой публикации. Можно задать номер или диапазон (больше, меньше, фиксированный диапазон).
- **Пакет** - задается в настройках проекта, осуществляется поиск по имени пакета.
- **Пользователь –** позволяет сортировать пакеты по имени пользователя, который запустил пакеты на обработку.
- **Статус** – отображается статус пакета (в обработке, завершено, запущен, не запущен, отвергнуто, отложено, перенесён). При статусе пакета «В обработке» над ним нельзя выполнять ни каких действий. Сначала должна быть завершена работа модуля.



Не запущен - документ просканирован системой, но его обработка не началась.

Завершено - документ ушел на экспорт, документ полностью доработан.

Запущен - пакет начал обработку. Используется для распознавания и экспорта – начал и пока не закончил обрабатываться, для валидации – пакет был открыт пользователем, но не был ни отложен, ни принят, ни отправлен на перераспознавание.

Отложено- пакет был открыт пользователем на валидации, а потом закрыт.

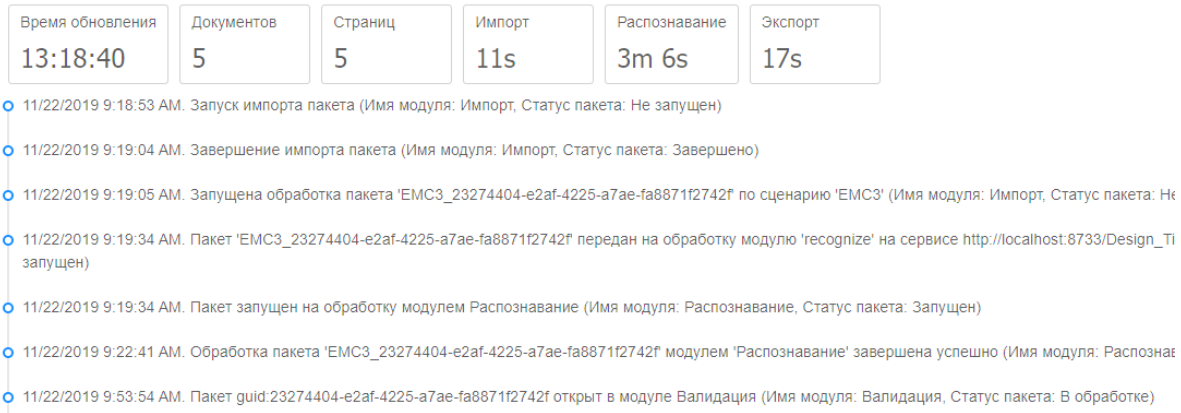
Отвергнуто- пакет завершил свою обработку с ошибкой.

Перенесён - при ручном перемещении между модулями.

Отправлен на доработку - статус, который принимает пакет в том случае, если в нем присутствует признак некорректности пакета и задано условие отправки таких документов по заданному электронному адресу.

- **Блокировка** - при работе какого-то модуля пакет блокируется и не доступен для изменения, также блокировка осуществляется при валидации пакета пользователем.
- **Модуль** – отображается модуль на котором находится пакет.

При открытии каждого пакета в отдельности можно увидеть все этапы его обработки.



Время обновления – время последнего обновления данных по выбранному пакету.

Документов – количество документов в выбранном пакете.

Страниц – количество страниц в выбранном пакете.

Импорт – время работы модуля импорта выбранного пакета.

Распознавание - время работы модуля распознавания выбранного пакета.

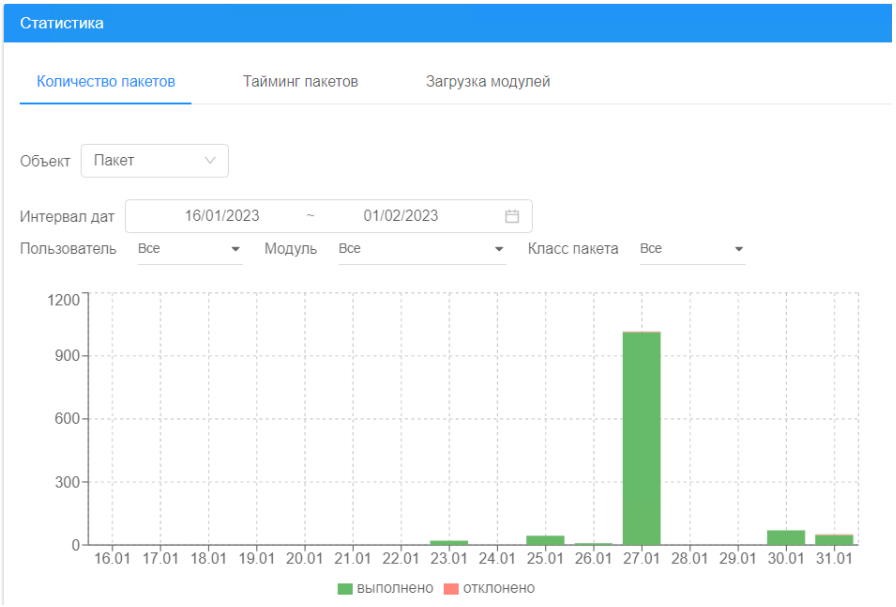
Экспорт – время работы модуля экспорт выбранного пакета.

В схеме работы модулей записывается точное время начала и окончания работы модулей. В схеме можно увидеть все основные этапы обработки пакета.

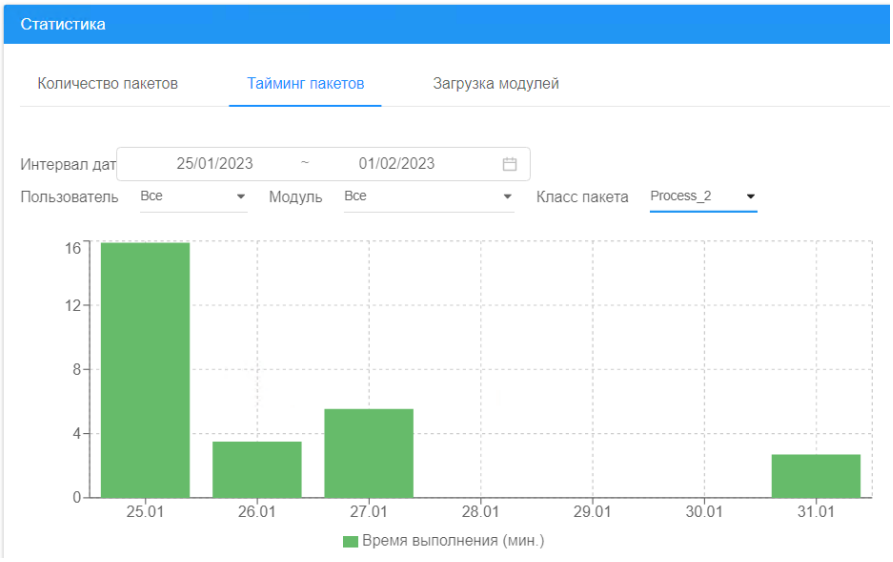
2.2.5 Статистика.

В статистике можно увидеть:

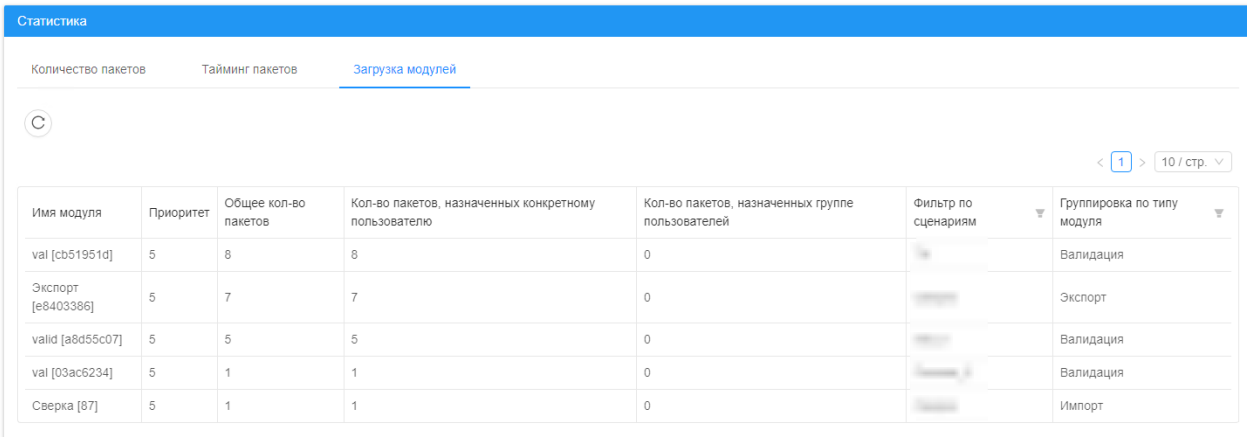
- Количество пакетов, обработанных за определенный промежуток времени.



- Тайминг пакетов – скорость обработки пакетов за определенный промежуток времени и определенными модулями.



- Загрузка модулей – общий отчет загрузки всех модулей системы по всем пакетам, с возможностью применения фильтров «Фильтр по сценариям» и «Группировка по типу модуля».



Имя модуля	Приоритет	Общее кол-во пакетов	Кол-во пакетов, назначенных конкретному пользователю	Кол-во пакетов, назначенных группе пользователей	Фильтр по сценариям	Группировка по типу модуля
val [cb51951d]	5	8	8	0		Валидация
Экспорт [e8403386]	5	7	7	0		Экспорт
valid [a8d55c07]	5	5	5	0		Валидация
val [03ac6234]	5	1	1	0		Валидация
Сверка [87]	5	1	1	0		Импорт

2.2.6 Внешние источники.

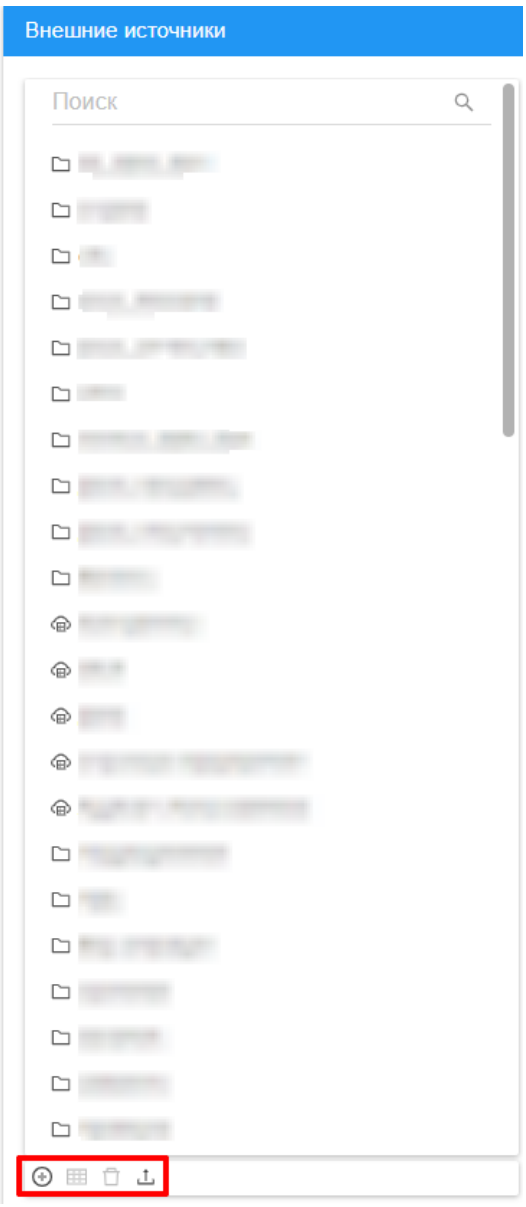
Под внешними источниками данных подразумеваются: Справочники, реляционные БД и выгрузки из них.

В качестве справочников могут использоваться любые данные:

- Наименования регионов
- Наименования городов
- Коды бухгалтерской отчетности
- Список имён и отчеств и др.

Так же возможно выполнить запрос во внешнюю систему обратившись к ее веб-сервису.

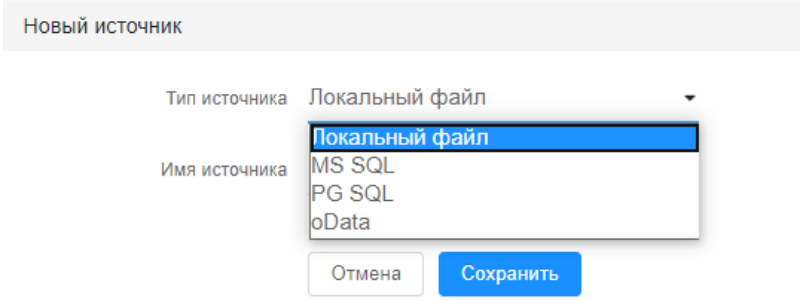
Инженер может создать свой внешний источник используя инструменты модуля администратора.



- Создание нового источника.
- Просмотр данных в выбранном источнике.
- Удалить выбранный источник.
- загрузить файл с данными в выбранный источник.

При создании нового внешнего источника необходимо:

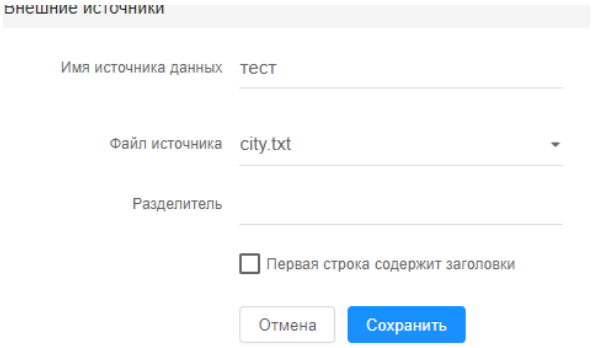
Выбрать тип внешнего источника и задать имя нового источника



Локальный файл ( ), База данных (MS SQL или PG SQL ( )), Внешний сервис (oData ( )).

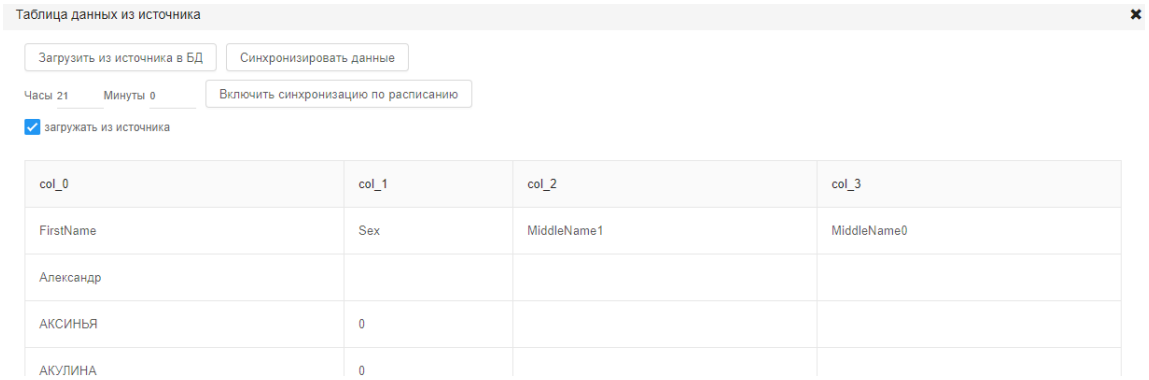
В зависимости от выбранного типа необходимо:

Тип «Локальный файл»



Выбрать файл источника и указать разделитель. Файл должен соответствовать указанному типу. При выборе файла со своего компьютера он будет сохранен на сервере в папке.

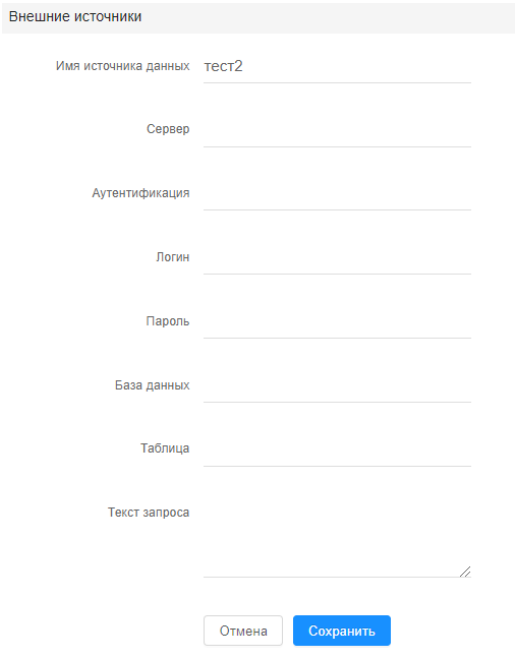
Нажав на кнопку «Данные из источника» можно просмотреть и провести настройку загружаемых данных:



В табличной области отображаются данные из источника в независимости загружены они в БД системы или нет. С помощью кнопок можно:

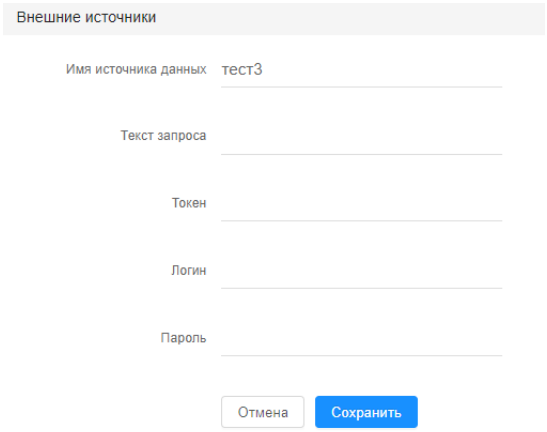
- **«Загрузить данные из источника в БД»** - данные из указанного источника попадают в БД системы;
- **«Синхронизировать данные»** - синхронизация данных из источника с данными в БД системы.
- **«Включить синхронизацию по расписанию»** - настройка синхронизации данных из источника с базой по расписанию. Указывается время сервера.

Тип «База данных» (MS SQL или PG SQL)

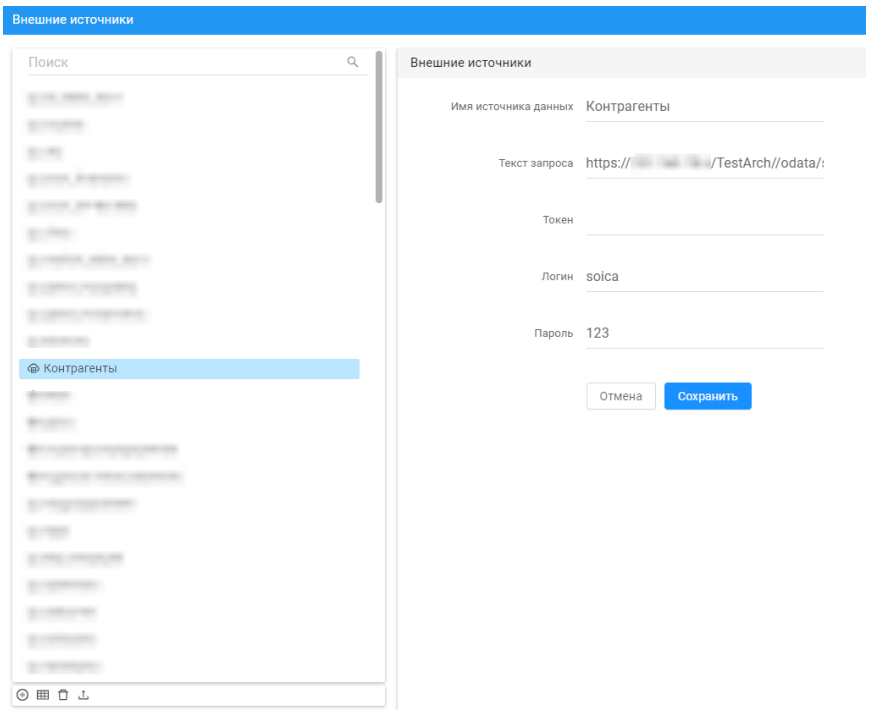


В настройках базы необходимо указать данные для аутентификации, таблицу с данными и запрос для выборки данных.

Тип «oData»



В настройках типа oData необходимо указать информацию о подключении к сервису, а также текст запроса.



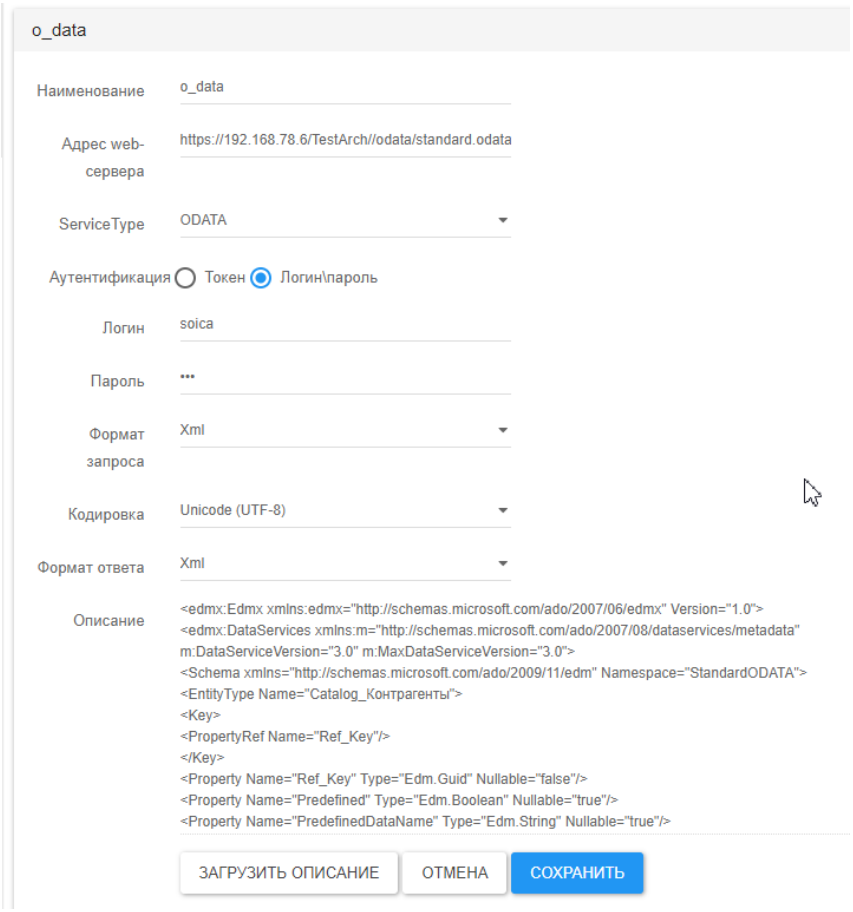
2.2.7 Веб сервисы

В системе предусмотрен инструмент обращения к сторонним веб сервисам для сравнения распознанных данных или выборки данных из имеющихся баз данных на веб сервисах.

Общий вид:

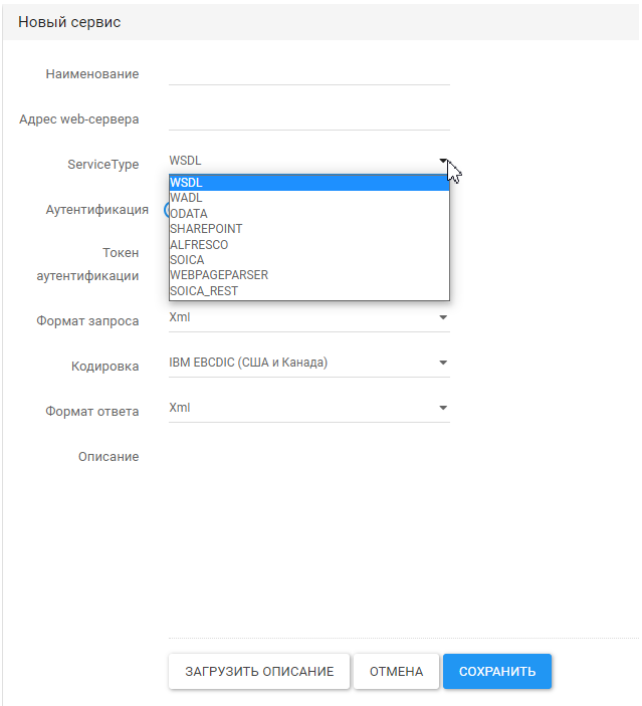
Для создания нового веб сервиса необходимо нажать на кнопку Добавить ( ). На общем виде представлен список доступных веб сервисов в системе. Веб сервисы не относятся к конкретному классу пакета или документа.

Меню Создание/Редактирование:



У каждого веб сервиса есть свое системное имя, по которому к нему можно будет обращаться при настройке проекта.

Типы аутентификации: по токену или логину/паролю. Поддерживаются следующие типы web-сервисов: WSLD, WADL, ODATA, SHAREPOINT, ALFRESCO, SOICA, WEBPAGEPARSER, SOICA_REST.



Параметры web-сервисов:

1. Наименование

2. Адрес описания web-сервиса. Для сервисов типа odata адрес должен содержать $metadata. Для сервисов wsdl и wadl это страница с описанием.

3. Тип сервиса

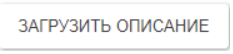
4. Тип аутентификации (при помощи токена или по логину-паролю)

5. Параметры аутентификации

6. Формат запроса (xml или json)

7. Кодировка запроса

8. Формат ответа (xml или json)

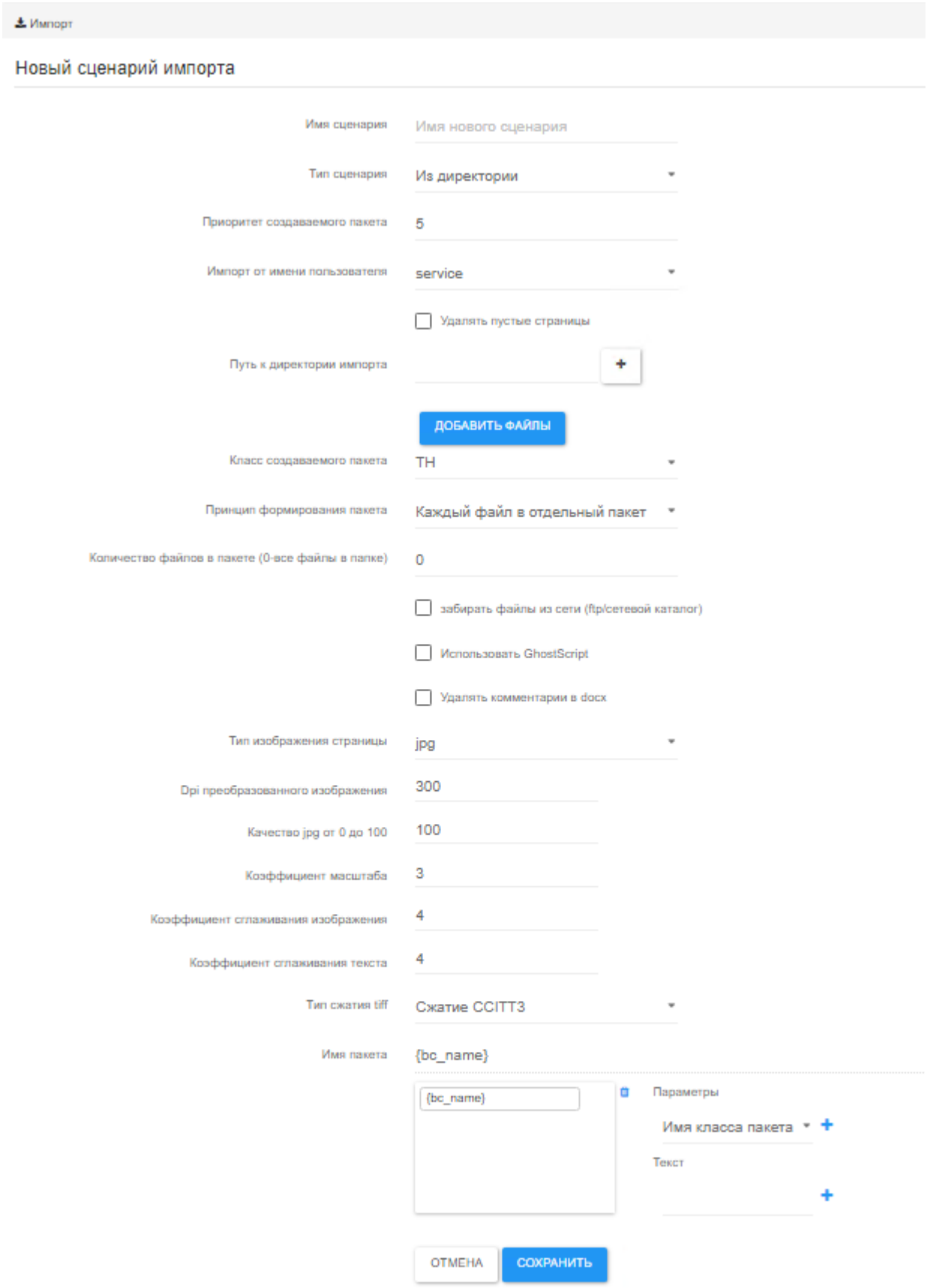
После сохранения настроек в поле описание автоматически появится ответ от сервера. Ответ от сервера можно получать принудительно кнопкой 

2.2.8 Способы обработки.

Способы обработки включают в себя 3 раздела:

- Автоимпорт.

В общем списке отображаются все созданные сценарии импорта. Для каждого класса пакета может быть выбрано несколько сценариев импорта.

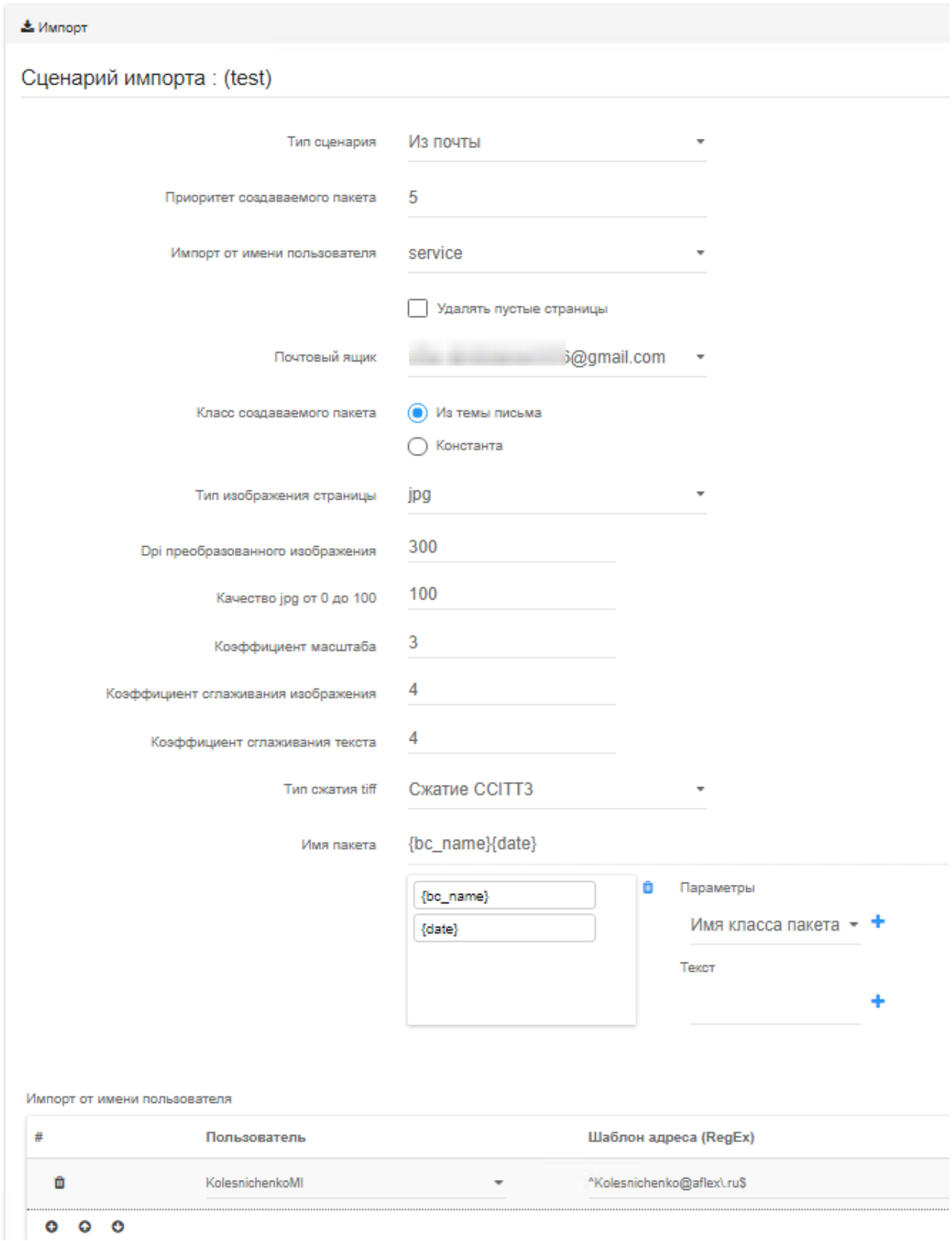


(Рис. 15. Настройка сценария импорта.)

При создании нового сценария необходимо:

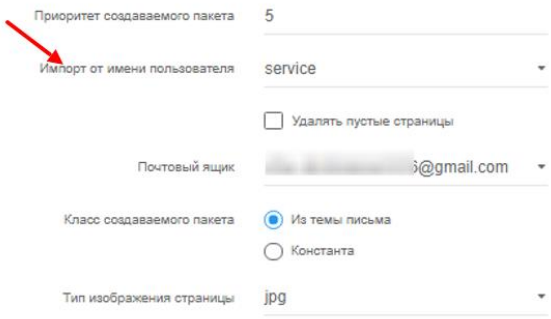
— Указать имя сценария;

— Выбрать Тип сценария. Существует два типа: Из директории и из почты;



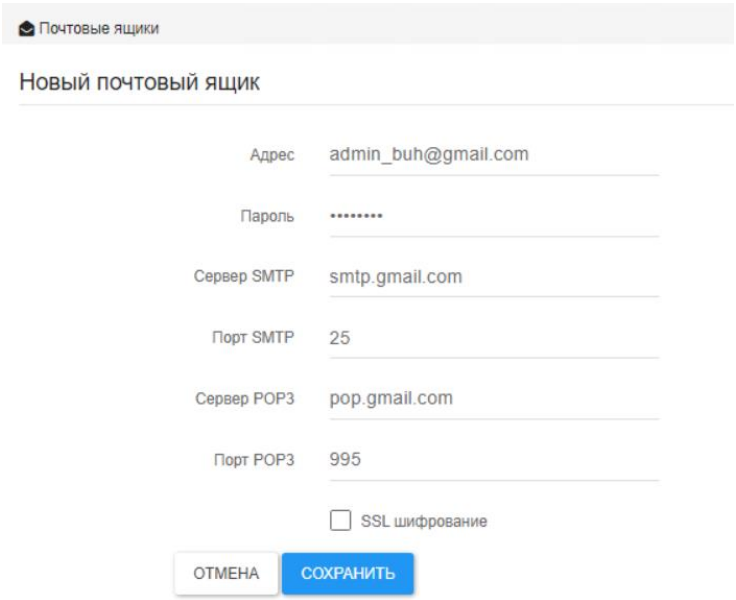
Если адрес почтового ящика отправителя письма соответствует регулярному выражению, указанному в столбце «Шаблон адреса», то импорт происходит от имени пользователя, к которому относится данное регулярное выражение.

Если адрес ящика не совпал ни с одним регулярным выражением, то импорт происходит от имени пользователя, указанного в строке «импорт от имени пользователя»



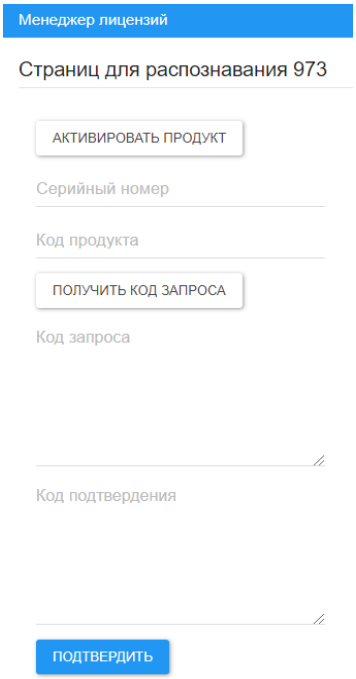
— Установить приоритет пакета. Чем выше приоритет пакета, тем быстрее он будет обрабатываться системой;
— Если выбран тип «Из директории», то необходимо указать путь до папки импорта. + - создать новую папку на сервере для импорта изображений.
— Добавлять файлы в сценарий импорта можно через кнопку ДОБАВИТЬ ФАЙЛЫ. Файлы будут добавлены в указанную выше папку на сервере. При включенном автоимпорте на сервере после добавления файлов сразу же начнется их обработка. Из папки импорта файлы удалятся, как только начинается обработка;
— Указать класс пакета по сценарию которого будут обрабатываться изображения.
— Принцип формирования пакета. Существует три принципа: 1. Каждый файл в отдельный пакет (пакет создается из каждого файла); 2. Несколько файлов в пакете. (в один пакет добавляется указанное ниже число файлов); 3. Содержимое подпапок. (в каждый отдельный пакет будет добавлено содержимое подпапок папки импорта).
— При установке отметки «забирать файлы с ftp» необходимо указать Путь к ftp каталогу, Логин и пароль. В остальном работа ведется как с обычной папкой.
— Тип изображения страницы – указывается в какой формат будет преобразованы страницы поступающего файла. Так же указываются характеристики обрабатываемого изображения: Dpi преобразованного изображения; Качество jpg (только для jpg, от 0 до 100); Тип сжатия tiff (только для tiff), коэффициент масштаба, коэффициент сглаживания изображения, коэффициент сглаживания текста.
— В импорте доступен конструктор имени пакета. В левый список необходимо добавлять параметры из выпадающего списка справа. Так же можно вручную ввести необходимый текст. Очередность параметров в левом списке определяет их последовательность в имени сформированного пакета.

- Счетчики.
  Счетчики можно применять в конструкторе полей пакета.
  В счетчике можно задать «шаг» от 1 до ∞. Затем счетчик можно использовать в присвоении имени пакету документа.
- Почтовые ящики.
  В этом разделе указываются электронные почтовые адреса для использования их в импорте или экспорте.



В настройках почтового ящика необходимо указать данные сервера, на котором располагается ящик. Это необходимо для связи системы с этим ящиком. Эти данные можно найти в описании почтового адреса на стороне сервера (gmail, yandex и т. д.)

2.2.9. Менеджер лицензий.

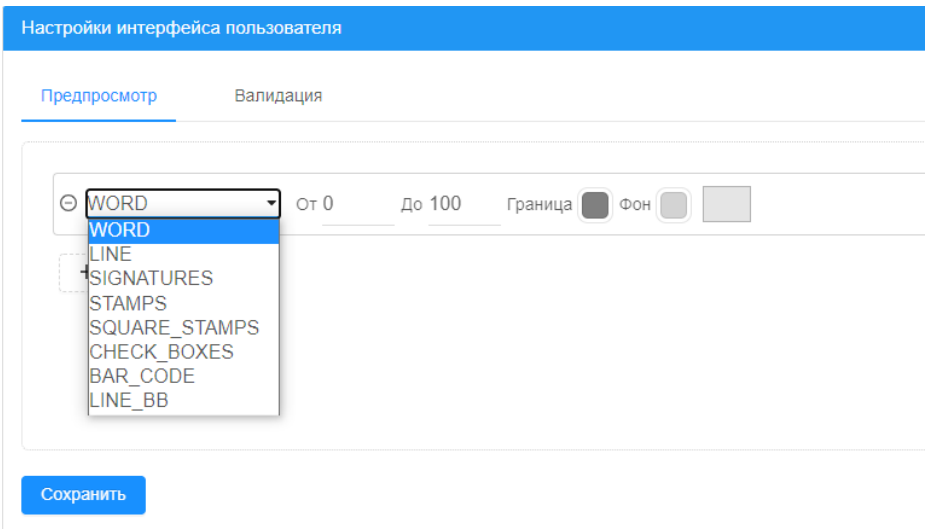


Лицензии – это количество страниц, которые могут поступить в систему на обработку. Перед распознаванием идет проверка на наличия необходимого количества лицензий для распознавания пакета, но сами лицензии списываются только после успешного завершения процесса распознавания.

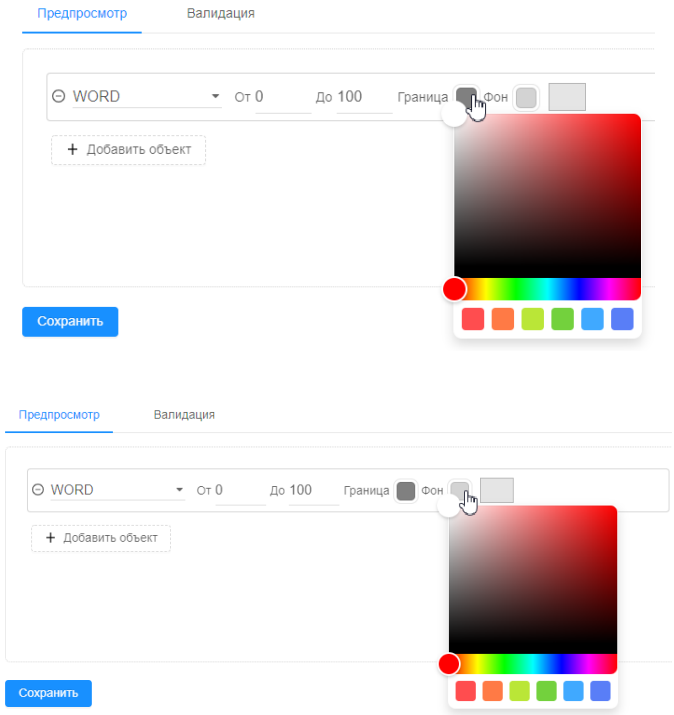
Для активации дополнительного числа лицензий необходимо иметь серийный номер и ключ.

2.2.10 Интерфейс пользователя.

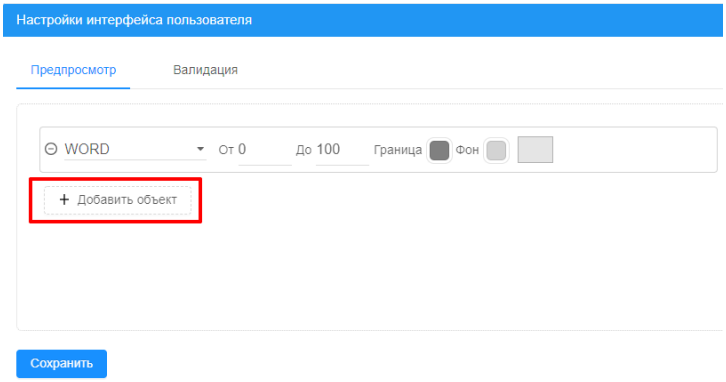
Этот раздел предназначен для настроек цвета отображения найденных элементов в результатах предпросмотра в модуле Администратора, а также для брендирования модуля Валидации.



Во вкладке Предпросмотр можно выбрать элемент из списка, выбрать для него цвет фона и границ, а также диапазон конфиденса, при котором данный элемент будет отображаться выбранным цветом в результатах предпросмотра.

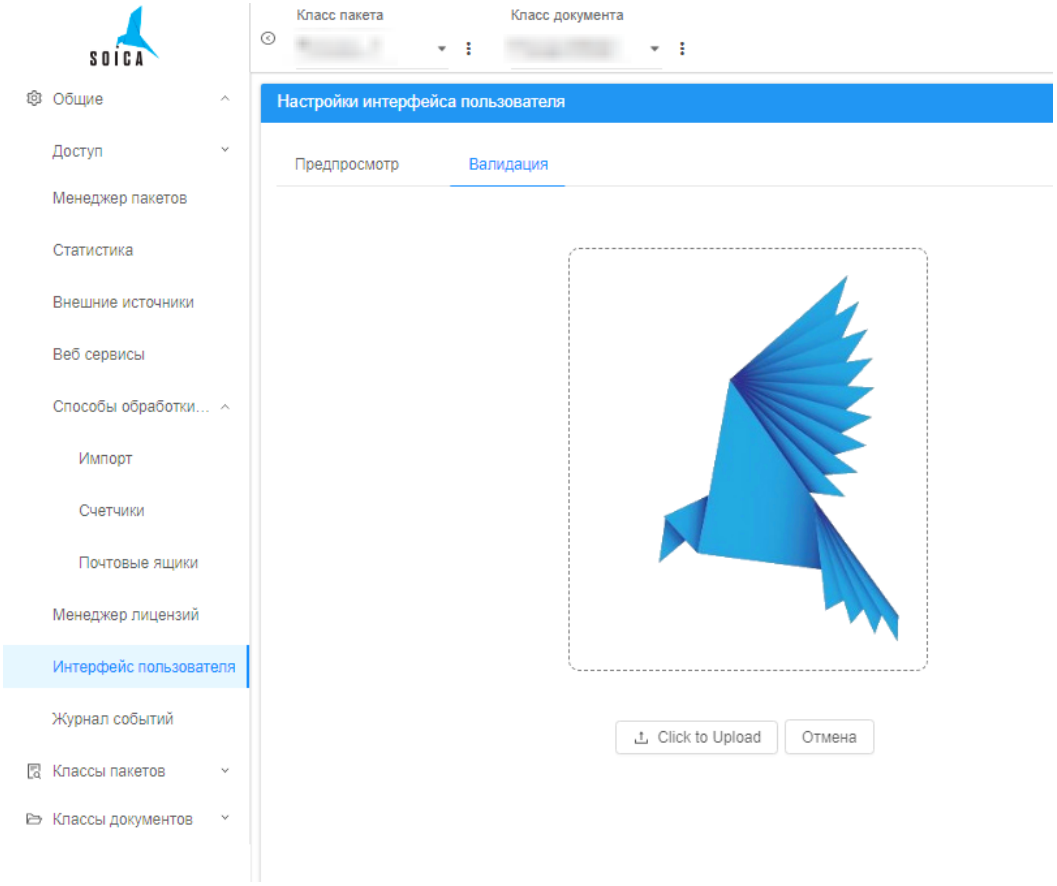


Для одного элемента можно задать несколько цветов (выбор цвета регулируется диапазоном конфиденса). Чтобы добавить новый элемент необходимо нажать кнопку Добавить объект.



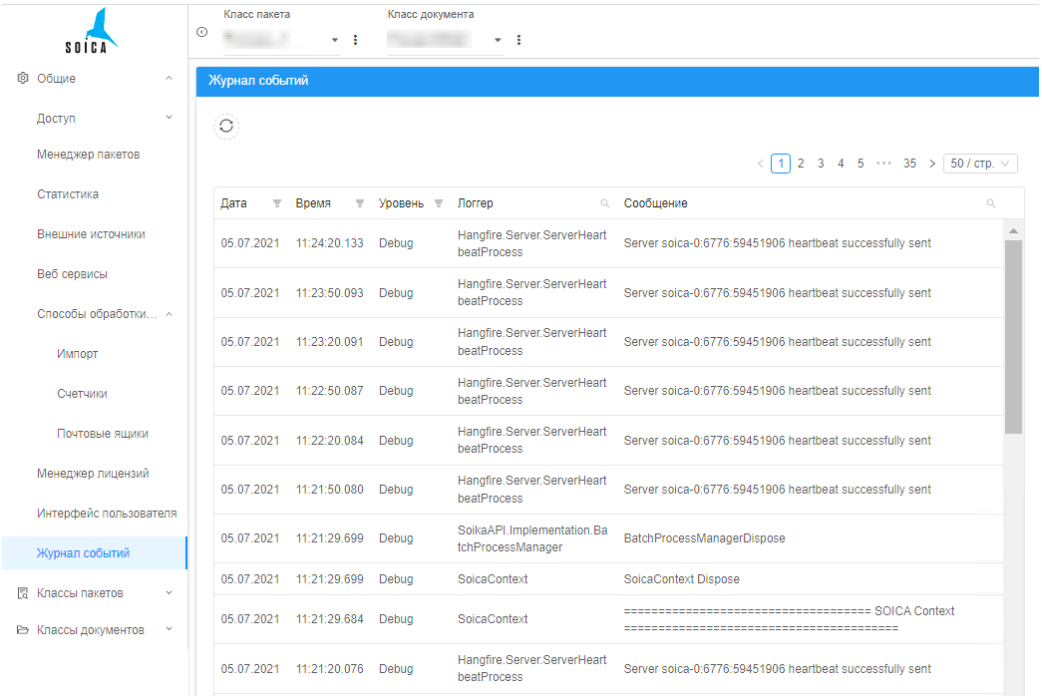
Количество создаваемых объектов неограниченно.

Во вкладке Валидация заказчик может загрузить свой фирменный логотип для отображения на странице ввода Логина и пароля в модуле валидации.



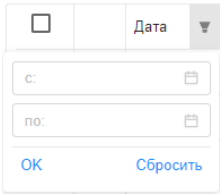
2.2.11 Журнал событий

Журнал событий необходим для отслеживая системных событий.

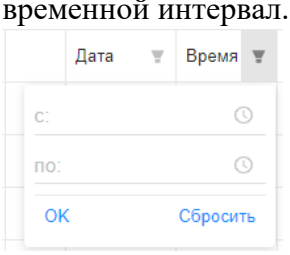


Основным элементом является список системных событий. В таблице мы видим:

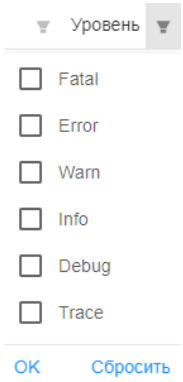
- **Дата** – столбец дата позволяет сортировать события за определенный период.



- **Время** - с помощью фильтра столбца «Время» можно задавать определенный временной интервал.



- **Уровень** - с помощью фильтра столбца «Уровень» можно задавать определенную категорию ошибок или системных сообщений.



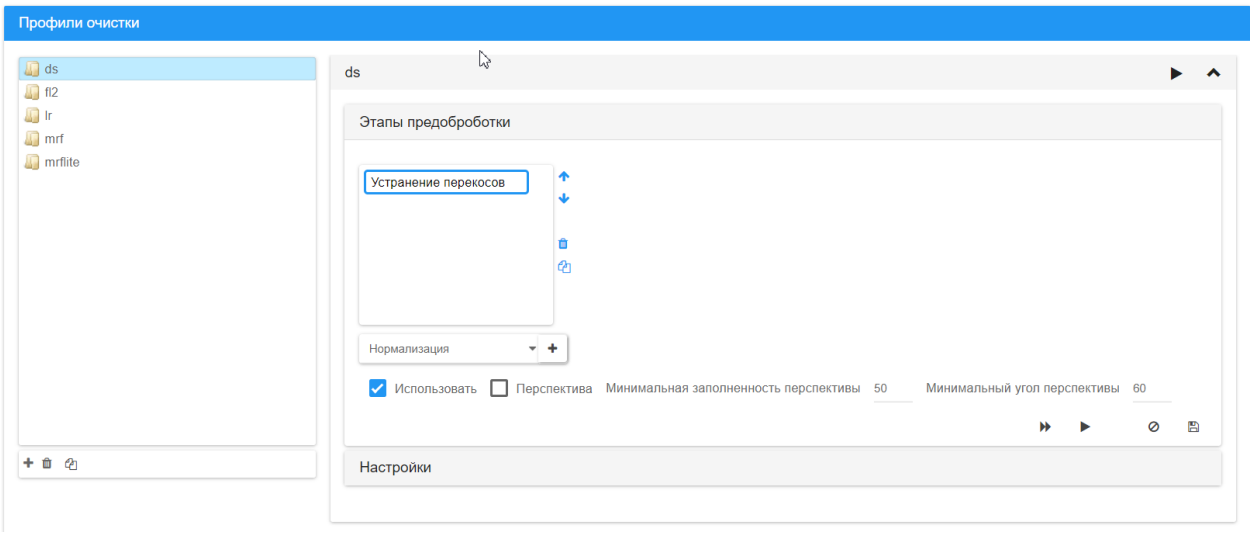
2.3 Меню распознавания (Классы пакетов).

В меню распознавания настраивается общий сценарий очистки изображения для лучшего поиска на нем OCR и настраиваются инструменты распознавания текста на изображениях.

2.3.1. Профили очистки.

Этап настройки профилей очистки называется **«Предобработка»**

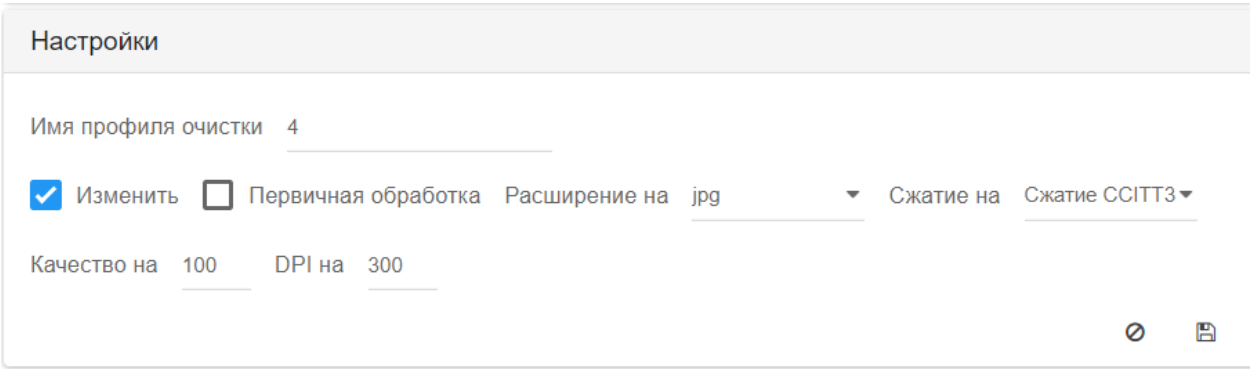
Цель предобработки: Хорошее качество распознавания требует качественного исходного изображения, что бывает не всегда. Предобработка позволяет скорректировать качество изображения так, чтобы необходимы данные можно было распознать.



(Рис. 16. Настройка профиля очистки)

Меню настройки профилей очистки состоит из:

— Список имеющихся профилей очистки. В этом списке можно создавать новый ( ), удалять ( ) или копировать ( ) профиль из другого класса пакета.
— Область настройки выбранного профиля. Область имеет два раздела: Этапы обработки и Настройки.



На форме общих настроек профиля есть:

Имя профиля очистки. Можно изменить имя профиля.

Изменить. Если данная опция выбрана, то в изображении репрезентации будут использоваться указанные параметры, а не параметры оригинального изображения

Первичная обработка. Если данная опция выбрана, то текущий профиль очистки будет применяться первым в очереди профилей очистки для каждого профиля распознавания. Только один профиль очистки может быть с этой опцией.

Расширение на. Расширение файла изображения репрезентации. Варианты: jpg, tif, png.

Сжатие на. Указывает тип сжатия для tif файла. Варианты: "Сжатие LZW", "Сжатие CCITT3", "Сжатие CCITT4", "Сжатие Rle", "Без сжатия".

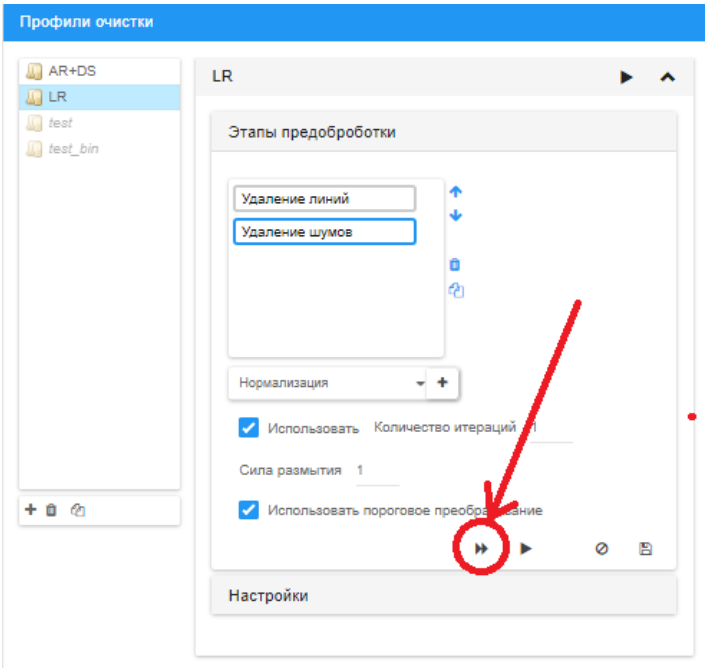
Качество на. Указывает качество сжатия для jpg файла. Диапазон: 0-100.

DPI на. Указывает количество точек на дюйм изображения репрезентации. Диапазон: 30-600.

На форме выбора этапов обработки из выпадающего списка необходимо выбрать нужный фильтр ( ) и провести его настройку, после этого добавить его в список фильтров нажав на кнопку «Сохранить» ( ).

Системой предусмотрен запуск всего профиля очистки к странице (- правый верхний угол), а также запуск отдельно этапов очистки (- правый нижний угол формы).

Кроме того, есть возможность запуска всех этапов из профиля до указанного этапа по очереди:



Каждый профиль очистки может состоять из набора разных графических фильтров.

Фильтры комбинируются, отрабатывают поочередно, учитывая результаты предыдущих. Поэтому важен порядок их расположения в списке фильтров.

Всего существует 18 фильтров.

1) Нормализация.

Нормализация – это выравнивание и обрезка лишней (пустой) информации. Используется для документов, удостоверяющих личность. Для коректной работы нормализации на документе должна быть фотография.

Общий принцип работы:

Первым опциональным этапом является устранение перекосов. Для которого с помощью границ Кенни и метода Хафа выполняется поиск доминирующих линий и поворот на вычисленный угол. Имеются настройки для данных функций.

Следующий необязательный этап – поиск корректной ориентации поворота. В этом этапе производится поиск объектов по каскадам Хаара на каждой из четырех ориентаций изображения. Где результат поиска лучше, та ориентация и является правильной. При отключенном этапе, выполняется поиск только в исходной ориентации.

Далее выполняется обрезка изображения по найденным по каскадам Хаара объектам, в соответствии с указанными настройками.

В системе есть натренированные каскады Хаара для поиска лица. Можно выполнять поиск по одному или нескольким каскадам.

Далее опционально выполняется устранение малых перекосов – этот этап аналогичен первому, но выполняется уже на обрезанном изображении.

Подбор контраста может выполняться для изменения яркости обрезанного изображения. Так же можно преобразовать изображение в оттенки серого. Так же можно выполнить бинаризацию изображения.

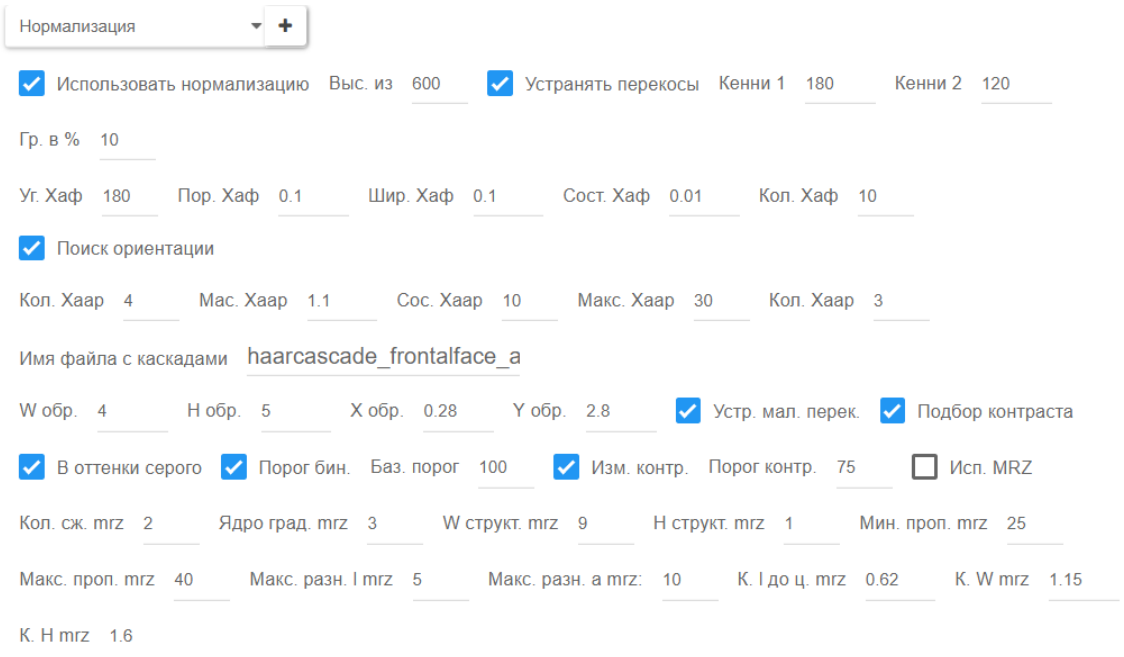
В случае не нахождения объекта каскадом можно выполнить изменение яркости изображения на указанное значение и направить его на повторный поиск.

ДУЛы с машиночитаемой зоной можно обрезать точнее, если использовать mrz обрезку. В этом случае выполняется поиск 2х строк машиночитаемой зоны по заданным параметрам. И далее выполняется корректирующая обрезка изображения.

(Рис. 17. Этапы работы фильтра: исходное изображение, границы Кенни, повернутое изображение с найденным лицом и границей обрезки, обрезанное изображение в оттенках серого)

(Рис. 18. Резлультаты работы mrz. Красная рамка – первая mrz строка, синяя – вторая mrz строка, фиолетовая – результат обрезки.)

Интерфейс настройки:



(Рис. 19 Настройка фильтра «Нормализация»)

Использовать нормализацию. Если эта опция выбрана, то к изображению будет применен этап предобработки – нормализация.

Выс. из. Указывает значение высоты изображения, к которому приводится оригинал перед дальнейшей обработкой. Чем меньше значение, тем быстрее будет выполняться обработка, но и тем больше вероятность получения некорректного результата.

Устранять перекосы. При выборе данной опции, на изображении будет произведен поиск линий, для последующего вычисление доминантного угла поворота линий и поворота изображения на этого угол. Алгоритм включает следующие параметры: Порог для Кенни 1, Порог для Кенни 2, Размер игнорируемой границы, Угловое разрешение Хафа, Порог Хафа, Длина Хафа, Соседи Хафа, Количество линий для подтверждения поворота:

- **Кенни 1**. Первый порог для преобразования Кенни. Диапазон: 0-255.
- **Кенни 2**. Второй порог для преобразования Кенни. Диапазон: 0-255.

Выделение границ Кенни использует два порога фильтрации: если значение пикселя выше верхней границы – он принимает максимальное значение (граница считается достоверной), если ниже – пиксель подавляется, точки со значением, попадающим в диапазон между порогов, принимают фиксированное среднее значение.

- **Размер игнорируемой границы.** Указывает размер границ изображения в процентах от его ширины и высоты, которые не будут участвовать в поиске линий. Диапазон: 1-50.
- **Угловое разрешение Хафа**. Разрешение для линии в бинарном изображении в градусах. Диапазон: 1-360.
- **Порог Хафа**. Минимальное количество пикселей в отрезке, выраженное в процентах от максимального размера изображения. Диапазон: 1-50.
- **Длина Хафа**. Минимальная толщина линии, выраженная в процентах от максимального размера изображения. Диапазон: 1-50.
- **Соседи Хафа**. Минимальное расстояние между линиями, выраженная в процентах от максимального размера изображения. Диапазон: 1-50.
- **Количество линий для подтверждения поворота**. Указывает минимальное количество найденных линий, при котором результаты поиска доминантного угла поворота признаются валидными.

Поиск ориентации. Если данная опция выбрана, то дальнейший поиск объектов по каскадам Хаара будет выполняться 4 раза, для каждой из 4х ориентаций изображения (поворот на угол кратный 90 градусов), ориентация выбирается по наилучшим результатам поиска по каскадам.

Поиск объекта по каскадам Хаара. Выполняет поиск объектов по каскадам Хаара на изображении, и выполняет обрезку изображения относительно результирующего объекта. Включает параметры: Количество каскадов, Фактор масштаба, Промежуток, Минимальный размер, Максимальный размер, Начало имени файла с каскадами, Регион обрезки (X, Y, Ширина, Высота):

- **Количество каскадов**. Максимальное количество файлов с каскадами, по которым будет производится поиск объектов на репрезентации. Диапазон значений: 1-4.
- **Фактор масштаба**. Параметр, показывающий на сколько будет меняться масштаб изображения, каждый проход поиска. Чем он меньше, тем дольше и подробнее будет выполняться поиск. Диапазон значений: 0,5-3.
- **Промежуток.** Параметр, определяющий, сколько соседей должен иметь каждый прямоугольник-кандидат. Этот параметр влияет на качество обнаруженных лиц. Более высокое значение приводит к меньшему количеству обнаружений, но с более высоким качеством.

- **Минимальный размер**. Указывает минимально возможный размер обнаружаемого объекта. Параметр выражается в процентах от ширины и высоты изображения. Диапазон значений: 1-100.
- **Максимальный размер**. Указывает максимально возможный размер обнаружаемого объекта. Параметр выражается в процентах от ширины и высоты изображения. Диапазон значений: 1-100.
- **Начало имени файла с каскадами.** Указывает начало наименования файлов с каскадами, которые будут использоваться для поиска объектов.
- **Регион обрезки.** Указывает область, которая будет вырезана из изображения:
  - **X.** Указывает отступ от результирующего объекта влево в долях от ширины результирующего объекта, где будет проходить левая граница вырезаемой области.
  - **Y.** Указывает отступ от результирующего объекта вверх в долях от высоты результирующего объекта, где будет проходить верхняя граница вырезаемой области.
  - **Ширина.** Указывает ширину вырезаемой области в долях от ширины результирующего объекта.
  - **Высота.** Указывает высоту вырезаемой области в долях от высоты результирующего объекта.

Устранение малых перекосов. Если выбрана эта опция, то после обрезки по каскадам Хаара, на результирующем изображении, будет выполнено устранение перекосов с настройками по умолчанию.

Подбор контраста. Если выбрана данная опция, то будет происходить бинаризация изображения с различным порогом до тех пор, пока верхняя половина изображения (за исключением 10-процентной границы) не станет «черной» на 5-10%, или пока не истечет количество итераций (20 штук).

В оттенки серого. Если выбрана данная опция, изображение переводится в оттенки серого.

Использовать бинаризацию. В текущей версии не используется.

Базовый порог бинаризации. Значение порога для бинаризации с которого начинается изменение контраста при подборе контраста. Диапазон: 0-255.

Изменение контраста. Если данная опция выбрана, то при неудачном поиске объекта по каскадам Хаара, будет выполнена бинаризация изображения с указанным порогом и повторно начата процедура поиска объектов.

Порог бинаризации при изменении контраста. Указывает значение порога для бинаризации, при изменении контраста для повторного поиска объектов по каскадам Хаара. Диапазон: 0-255.

Использовать MRZ. Если выбрана данная опция, то будет производится поиск 2хстроковой машиночитаемой зоны, для последующего поворота и обрезки по ней. Имеющиеся настройки: Количество сжатий, Ядро градиента, Ширина структуры, Высота структуры, Минимальная пропорция, Максимальная пропорция, Максимальная разница длин строк, Максимальный угол между строками, Множитель расстояния до центра, Множитель ширины, Множитель высоты:

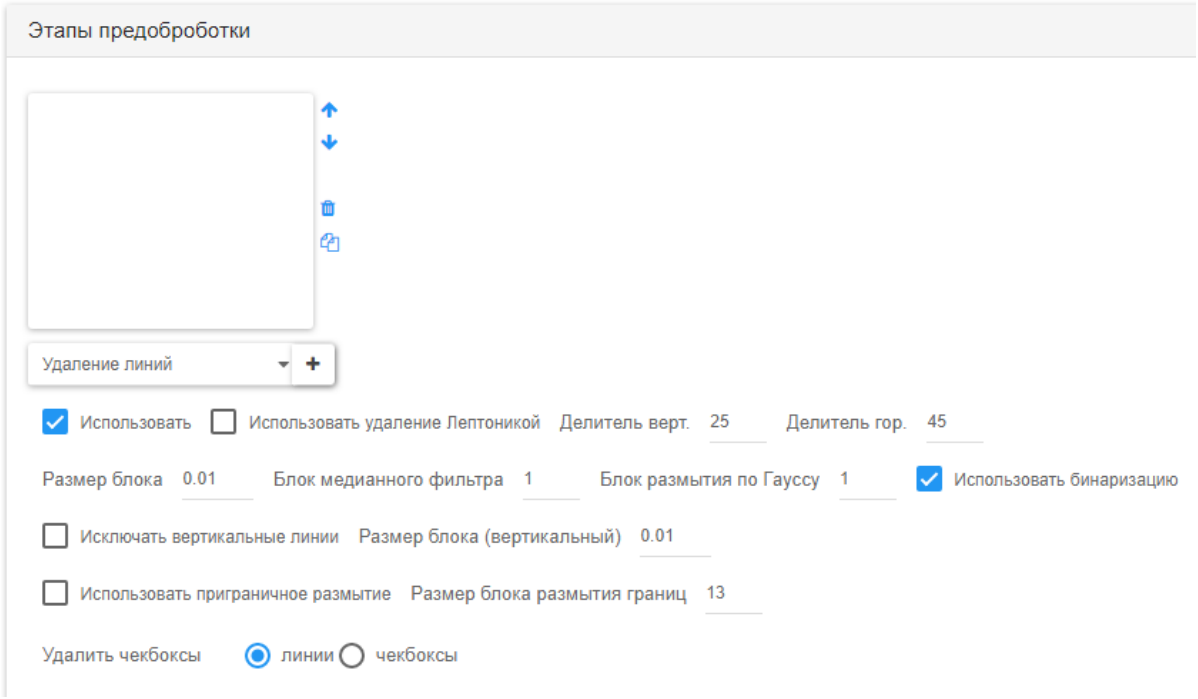
- **Количество сжатий.** Число, указывающее сколько раз изображение будет уменьшено вдвое перед поиском строк машиночитаемой зоны. Диапазон: 0-10.

- **Ядро градиента.** Указывает размер эллипса для морфологического преобразования типом градиента. Чем больше величина, тем сильнее будут объединяться объекты на изображении. Диапазон: 3-15 (с шагом 2).
- **Ширина структуры.** Указывает ширину структуры для морфологического поиска объекта, в долях от его высоты. Диапазон: 1-150.
- **Высота структуры.** Указывает высоту структуры для морфологического поиска объекта, в долях от его ширины. Диапазон: 1-150.
- Ширина и высота структуры указывают пропорции искомого объекта.
- **Минимальная пропорция**. Указывает минимальное соотношение ширины и высоты прямоугольника, описанного вокруг найденного контура, для того чтобы он мог считаться строкой машиночитаемой зоны. Диапазон: 1-100.
- **Максимальная пропорция**. Указывает максимальное соотношение ширины и высоты прямоугольника, описанного вокруг найденного контура, для того чтобы он мог считаться строкой машиночитаемой зоны. Диапазон: 1-100.
- **Максимальная разница длин строк**. Указывает максимальное значение в процентах от длины первой строки машиночитаемой зоны, на которое может отклоняться длина второй машиночитаемой зоны, чтобы считаться строками машиночитаемой зоны. Диапазон: 0-30.
- **Максимальный угол между строками**. Указывает максимальное значение угла между строками машиночитаемой зоны, чтобы считать эти строки строками машиночитаемой зоны. Диапазон: 0-30.
- **Множитель расстояния до центра**. Указывает расстояние от верхней строки машиночитаемой зоны до центра итогового изображения, в долях от ширины этой строки. Диапазон: 0,3-1 (шаг 0,01).
- **Множитель ширины**. Указывает ширину итогового изображения, в долях от ширины верхней строки машиночитаемой зоны. Диапазон: 1-2 (шаг 0,01).
- **Множитель высоты**. Указывает высоту итогового изображения, в долях от ширины верхней строки машиночитаемой зоны. Диапазон: 1-2 (шаг 0,01).

2) Удаление линий.

Этап предобработки служит для удаления вертикальных и/или горизонтальных линий на изображении, а так же чекбоксов.

Интерфейс настройки



(Рис. 20 Настройка фильтра «Удаление линий»)

Использовать. Если эта опция выбрана, то к изображению будет применен этап предобработки – удаление линий.

Использовать удаление Лептоникой. При выборе этого пункта будут использоваться библиотеки Лептоника для нахождения линий. Все операции с матрицей изображения (обработка граней, бинаризация и т д) будут такие же.

Делитель верт. Указывает число, на которое делится высота изображения для получения высоты структуры для морфологического поиска вертикальных линии. Диапазон: 1-300.

Делитель гор.. Указывает число, на которое делится ширина изображения для получения ширины структуры для морфологического поиска горизонтальных линии. Диапазон: 1-300.

Размер блока. Указывает размер блока для адаптивной бинаризации, выраженный в процентах от диагонали изображения. Диапазон: 0-100.

Блок медианного фильтра. Указывает размер блока для медианного фильтра. Диапазон: 1-25 (шаг 2).

Блок размытия по Гауссу. Указывает размер блока для фильтра по Гауссу. Диапазон: 1-25 (шаг 2).

Использовать бинаризацию. Опция указывает будет ли бинаризировано исходное изображение перед поиском линий.

Исключать вертикальные линии. Данная опция позволяет использовать щадящий режим удаление горизонтальных линий, в том случае если эти линии пересекают текст.

Размер блока (вертикальный). Указывает размер блока для адаптивной бинаризации, выраженный в процентах от диагонали изображения используемый для

поиска вертикальных линий при выбранной опции «Исключать вертикальные линии». Диапазон: 0-100.

Использовать приграничное размытие. Дополнительная очистка изображения от мелких деталей около линий.

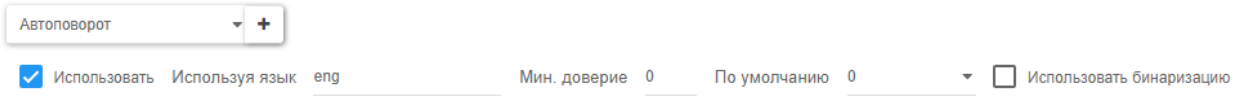
Размер блока размытия границ. Размер блока определяет ширину прямоугольников с обеих сторон от линии, к которым будет применено размытие.

Удалять чекбоксы. При выборе данной опции на изображении будет проведен автоматический поиск и удаление чекбоксов.

3) Автоповорот.

Данный этап предобработки выполняет автоповорот изображения на угол кратный 90 градусов по тексту. Поворот производится относительно найденного текста.

Интерфейс настройки



(Рис. 21 Настройка фильтра «Автоповорот»)

Использовать. Если эта опция выбрана, то к изображению будет применен этап предобработки – автоповорот.

Используя язык. Указывает на буквенное обозначение языка (или нескольких языков через символ «+»), используя которые будет выполнен автоповорот. Если языки не указаны, информация будет взята из настроек профиля распознавания.

Минимальный процент доверия. Указывает минимальный процент доверия к результату поиска корректной ориентации изображения, при котором будет выполнен автоповорот.

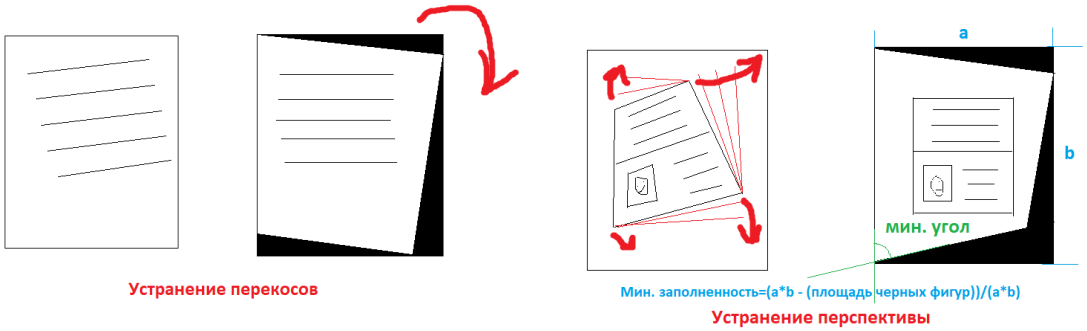
По умолчанию. Указывает угол, на который будет повернуто изображение, если степень доверия к результату поиска корректной ориентации ниже чем указанная. Возможные значения: 0, 90, 180, 270.

Использовать бинаризацию. При выборе данной опции перед поворотом изображения к нему будет применена бинаризация. После автоповорота изображение будет переведено на следующий этап без бинаризации.

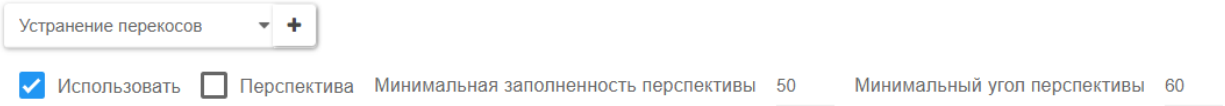
4) Устранение перекосов.

Устранение углового перекоса изображений и выравнивание текста на изображении.

Общий принцип работы:



Интерфейс настройки



(Рис. 22 Настройка фильтра «Устранение перекосов»)

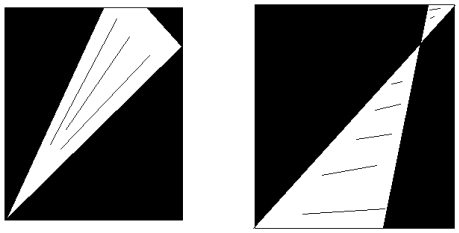
Использовать. Если эта опция выбрана, то к изображе нию будет применен этап предобработки – устранение перекосов.

Перспектива. Область, оставшаяся по краям после выравнивания изображения.

Минимальная заполненность перспективы. Минимально угол четырехугольника, в который трансформируется исходный прямоугольник, для того чтобы данное устранение перспективы считалось не ошибочным.

Минимальный угол перспективы. Минимальный процент исходного изображения на итоговом изображении, для того чтобы данное устранение перспективы считалось не ошибочным.

Иногда без указания этих настроек устранение перспективы отрабатывает не верно.

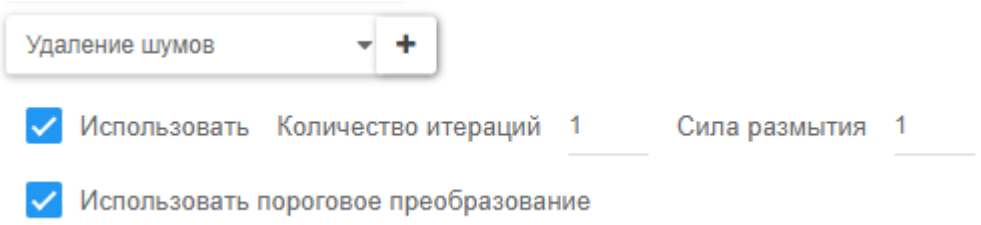


Примеры ошибок устранения перспективы

5) Удаление шумов.

Данный этап предобработки выполняет последовательное уменьшение и увеличение изображения с целью уничтожения шума. Уничтожение шума добивается за счет «размытия» изображения. За одну итерацию изображение уменьшается в два раза и увеличивается в два раза.

Интерфейс настройки.



(Рис.23 Настройка фильтра «Удаление шумов»)

Использовать. Если эта опция выбрана, то к изображению будет применен этап предобработки – удаление шумов.

Количество итерация. Указывает количество попеременных уменьшений и увеличений изображения. Чем больше, тем более «размытое» будет изображение. Диапазон: 0-10.

Сила размытия – во сколько раз будет происходить уменьшение изображение перед увеличением

Пороговое преобразование – уменьшать ли количество цветов в результате до 8ми битного.

6) Бинаризация.

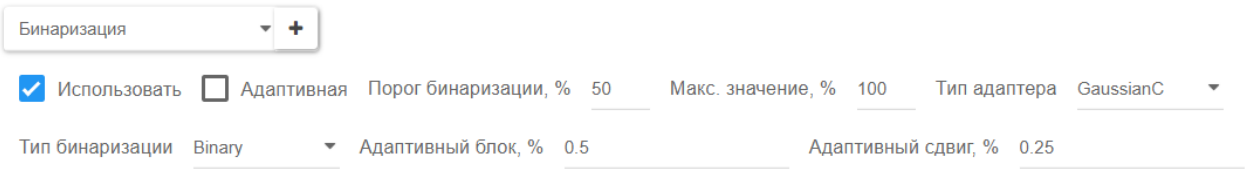
Данный этап предобработки делает изображение монохромным в соответствии с указанными параметрами. Т. е. переводит исходное изображение в двухцветное черно-белое.

Общий принцип работы.

Подробнее: https://docs.opencv.org/3.4/d7/d4d/tutorial_py_thresholding.html

(Рис. 24. Этапы работы фильтра)

Интерфейс настройки.



(Рис. 25 Настройка фильтра «Бинаризация»)

Использовать. Если эта опция выбрана, то к изображению будет применен этап предобработки – бинаризация.

Адаптивна. Если выбрана эта опция, то порог бинаризации будет вычисляться в каждой области изображения (с указанным размером).

Порог бинаризации. Величина, указывающая величину яркости пикселя, выше которой он становится белым, если ниже – черным (для типа бинаризации Binary). Диапазон выражен в процентах: 0-100.

Максимальное значение. Указывает значение яркости пикселя в процентах, используемое в типах бинаризации: Binary, BinaryInv. Диапазон: 0-100.

Тип адаптера. Указывает алгоритм вычисления порога в пределах указанного блока изображения. Варианты: MEAN_C, GAUSSIAN_C.

- **MEAN_C**. Пороговое значение представляет собой среднее значение размера блока × размер окрестности минус значение сдвига.
- **GAUSSIAN_C**. Пороговое значение является взвешенной суммой (кросс-корреляции с Гауссовым окном) из блока × размер окрестности минус значение сдвига. Значение сдвига по умолчанию (стандартное отклонение) используется для указанного блока.

Тип бинаризации. Указывает на то как будет изменяться пиксели изображения при достижении или не достижении указанного порога. Варианты: Binary, BinaryInv, Trunc, Tozero, TozeroInv, Mask, Otsu, Triangle:

- **Binary.** Пиксель закрашивается указанным максимальным значением, если яркость его выше порога, иначе – черным.
- **BinaryInv**. Пиксель закрашивается черным, если яркость его выше порога, иначе – закрашивается указанным максимальным значением.
- **Trunc**. Пиксель заполняется цветом с яркостью порога, если яркость его выше порога, иначе – не меняется.
- **Tozero**. Пиксель не меняется, если яркость его выше порога, иначе – черным.
- **TozeroInv**. Пиксель закрашивается черным, если яркость его выше порога, иначе – не меняется.
- **Mask.** Не реализуется в данном этапе.
- **Otsu.** Не реализуется в данном этапе.
- **Triangle.** Не реализуется в данном этапе.

Адаптивный блок. Указывает размер блока для расчета порога бинаризации, в процентах от диагонали изображения. Диапазон: 0-100 (шаг 0,01).

Адаптивный сдвиг. Указывает величину, на которую смещается значение порога при адаптивной бинаризации, в процентах от диагонали изображения. Диапазон: 0-100 (шаг 0,01).

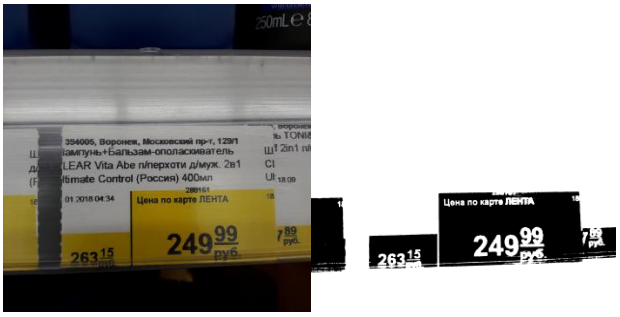
7) Фильтр по компонентам HSV

Данный этап раскладывает изображение по каналам цветовой модели. Используется для выделения какой-либо цветовой области или объекта. Настройки подбираются под определенный тип цветовой модели.

Общий принцип работы.

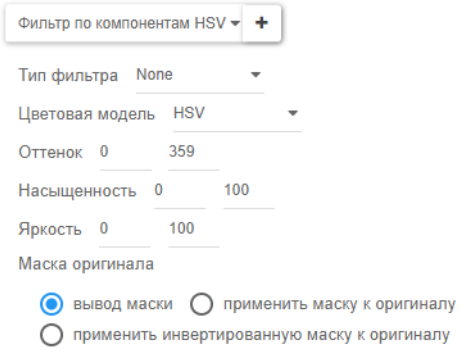
Цветовые модели HSV (Оттенок, Насыщенность и Яркость)

Далее выполняет пороговое преобразование по каждому из каналов по указанному максимальному и минимальному значениям и складывает (логическое сложение) получившиеся изображения.



(Рис. 26 Этапы работы фильтра)

Интерфейс настройки.



(Рис. 27 Настройка фильтра «Фильтр по компонентам HSV»)

Тип фильтра. Указывает цветовое пространство, компоненты которого будут фильтроваться. Варианты: HSV, V, S, SV, H, HV, HS.

Цветовая модель. Варианты: HLS, LAB, YCrCb, RGB.

Оттенок. Если выбрана эта опция, то отфильтрованный канал оттенка будет использоваться в качестве слагаемого для результирующего изображения.

— **Минимальное значение оттенка**. Указывает значение оттенка пикселя изображения, ниже которого пиксель закрасится белым. Диапазон: 0-359.
— **Максимальное значение оттенка**. Указывает значение оттенка пикселя изображения, выше которого пиксель закрасится белым. Диапазон: 0-359.

Пиксель со значением больше минимального и меньше максимального закрашивается черным цветом. Минимальное значение не может быть больше максимального.

Насыщенность. Если выбрана эта опция, то отфильтрованный канал насыщенности будет использоваться в качестве слагаемого для результирующего изображения.

— **Минимальное значение насыщенности**. Указывает значение насыщенности пикселя изображения, ниже которого пиксель закрасится белым. Диапазон: 0-100.

— **Максимальное значение насыщенности**. Указывает значение насыщенности пикселя изображения, выше которого пиксель закрасится белым. Диапазон: 0-100.

Пиксель со значением больше минимального и меньше максимального закрашивается черным цветом. Минимальное значение не может быть больше максимального.

Яркость. Если выбрана эта опция, то отфильтрованный канал яркости будет использоваться в качестве слагаемого для результирующего изображения.

— **Минимальное значение яркости**. Указывает значение яркости пикселя изображения, ниже которого пиксель закрасится белым. Диапазон: 0-100.

— **Максимальное значение яркости**. Указывает значение яркости пикселя изображения, выше которого пиксель закрасится белым. Диапазон: 0-100.

Вывод маски – выводит отсеянные пиксели черным цветом, остальные – белым.

Применять маску к оригиналу – те пиксели, которые были черными с первой опцией, будут оригинальными, остальные – белыми.

Применять инвертированную маску к оригиналу – те пиксели, которые были черными с первой опцией, будут белыми, остальные – оригинальными

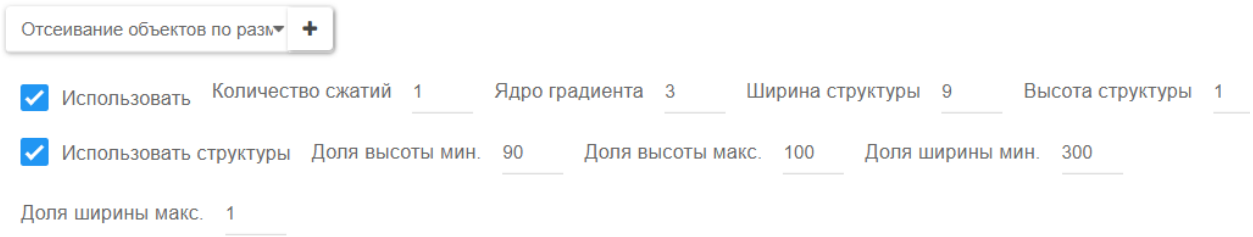
Пиксель со значением больше минимального и меньше максимального закрашивается черным цветом. Минимальное значение не может быть больше максимального.

8) Отсеивание объектов по размеру.

Данный этап предобработки выполняет поиск контуров, приведение их к прямоугольникам, отсеивание прямоугольников по указанным диапазонам размеров, и наложение полученной из отсеянных прямоугольников маски на изображение.

Условно говоря на изображении остаются только те объекты (текст, изображения, символы и т. д.), которые подходят под заданные параметры максимальной и минимальной высоты и ширины относительно параметров страницы.

Интерфейс настройки.



(Рис. 28. Настройка фильтра «Отсеивание объектов по размеру»)

Использовать. Если эта опция выбрана, то к изображению будет применен этап предобработки – отсеивание объектов по размеру.

Количество сжатий. Число, указывающее сколько раз изображение будет уменьшено вдвое перед поиском контуров. Диапазон: 0-10.

Ядро градиента. Указывает размер эллипса для морфологического преобразования типом градиента. Чем больше величина, тем сильнее будут объединяться объекты на изображении. Диапазон: 3-15 (с шагом 2).

Ширина структуры. Указывает ширину структуры для морфологического поиска объекта, в долях от его высоты. Диапазон: 1-150.

Высота структуры. Указывает высоту структуры для морфологического поиска объекта, в долях от его ширины. Диапазон: 1-150.

Использовать структуру. Данная опция указывает будет ли выполняться поиск контуров на изображении после морфологических преобразований:

Доля высоты мин. Указывает минимальную высоту прямоугольника, который попадет в маску в процентах от высоты изображения. Диапазон: 0.00-100.

Доля высоты макс. Указывает максимальную высоту прямоугольника, который попадет в маску в процентах от высоты изображения. Диапазон: 0.00-100.

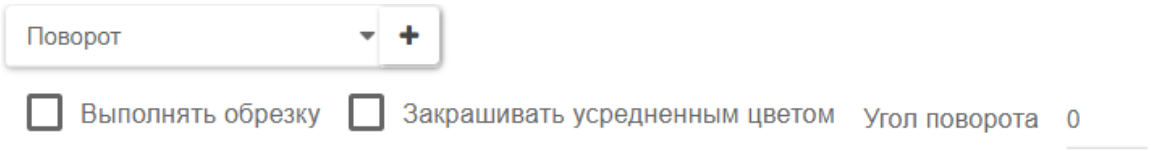
Доля ширины мин. Указывает минимальную ширину прямоугольника, который попадет в маску в процентах от ширины изображения. Диапазон: 0.00-100.

Доля ширины макс. Указывает максимальную ширину прямоугольника, который попадет в маску в процентах от ширины изображения. Диапазон: 0.00-100.

9) Поворот.

Данный этап предобработки выполняет поворот изображения на заданный угол. В отличии от автоповорота, данный этап поворачивает изображение в независимости от расположения данных на изображении. Используется только в том случае, если заранее известно, что все поступающие документы на обработку будут повернуты одинаково.

Интерфейс настройки



(Рис. 29 Настройка фильтра «Поворот»)

Выполнять обрезку. Данная опция определяет оставлять ли размер изображения после поворота исходным или увеличивать его, для полного охвата повернутого изображения.

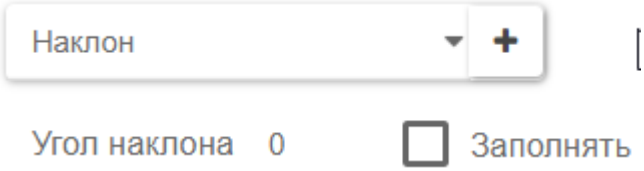
Закрашивать усредненным цветом. Если эта опция выбрана, то области, образуемые в результате поворота на непрямой угол, будут закрашены цветом, полученный из исходного изображения, иначе – черным.

Угол. Указывает угол, на который будет повернуто изображение. Диапазон значений: 0-359.

10) Наклон.

Данный этап предобработки выполняет наклон (построчный горизонтальный сдвиг) изображения на указанный угол. В отличии от устранения перекосов, данный этап выполняет наклон изображение в независимости от расположения данных на изображении.

Интерфейс настройки.



(Рис. 30 Настройка фильтра «Наклон»)

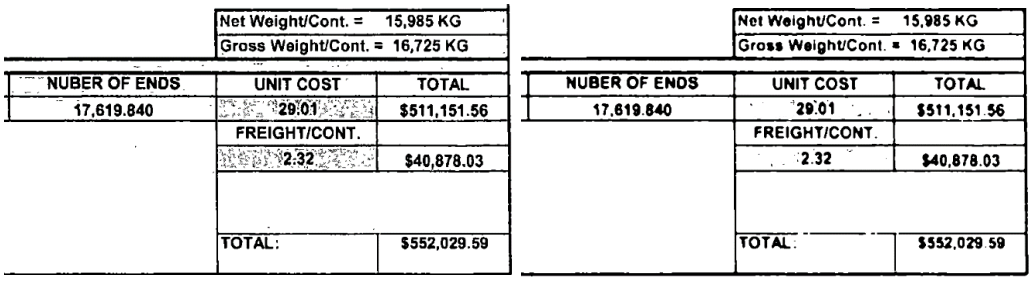
Угол наклона. Указывает угол, на который будет наклонено изображение. Диапазон: -45 - 45.

Заполнять. При выборе данной опции, области, образуемые при наклоне изображения заполняются рассчитанным усредненным цветом.

11) Медианный фильтр.

Данный этап предобработки выполняет медианное размытие с указанным ядром. Размером ядра корректируется степень размытия изображения.

Пример использования:

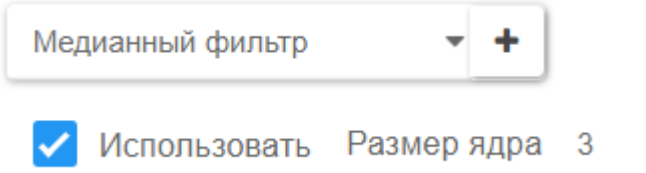


	Net Weight/Cont. = 15,985 KG	
	Gross Weight/Cont. = 16,725 KG	
NUBER OF ENDS	UNIT COST	TOTAL
17,619.840	29.01	$511,151.56
	FREIGHT/CONT.	
	2.32	$40,878.03
	TOTAL:	$552,029.59

	Net Weight/Cont. = 15,985 KG	
	Gross Weight/Cont. = 16,725 KG	
NUBER OF ENDS	UNIT COST	TOTAL
17,619.840	29.01	$511,151.56
	FREIGHT/CONT.	
	2.32	$40,878.03
	TOTAL:	$552,029.59

(Рис. 31 Пример использования фильтра «Медианный фильтр»)

Интерфейс настройки



(Рис. 32 Настройка фильтра «Медианный фильтр»)

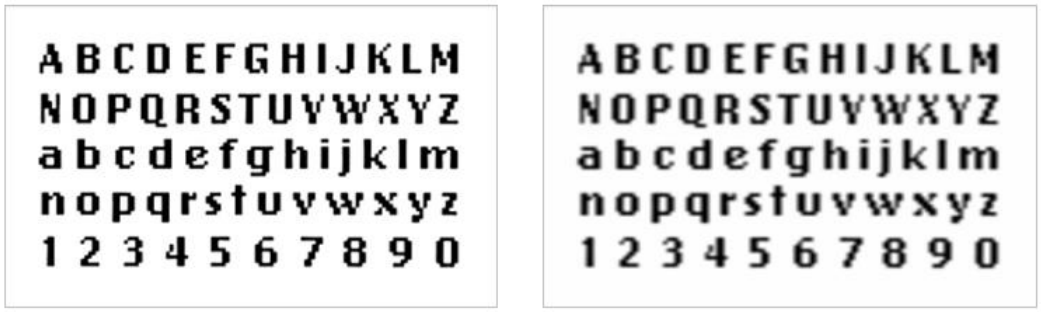
Использовать. Если эта опция выбрана, то к изображению будет применен этап предобработки – медианный фильтр.

Размер ядра. Указывает размер ядра медианного размытия. Диапазон: 1-25 (шаг 2). Чем больше размер ядра, тем сильнее будет размытие.

12) Размытие по Гауссу.

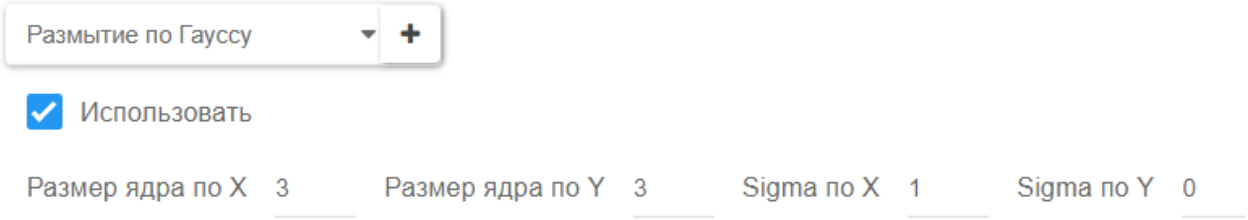
Данный этап предобработки выполняет размытие по Гауссу с указанными настройками. Здесь можно более тонко настроить размытие относительно ядра по осям X и Y.

Пример использования:



(Рис. 33 Пример использования фильтра «Размытие по Гауссу)

Интерфейс настройки.



(Рис. 34 Настройка фильтра «Размытие по Гауссу)

Использовать. Если эта опция выбрана, то к изображению будет применен этап предобработки – Гаусс фильтр.

Ядро X. Размер ядра размытия по оси X. Диапазон: 1-25 (шаг 2).

Ядро Y. Размер ядра размытия по оси Y. Диапазон: 1-25 (шаг 2).

Sigma X. Стандартное отклонение гауссова ядра в направлении X. Диапазон: 1-100.

Sigma Y. Стандартное отклонение гауссова ядра в направлении Y, если оно 0, то значение берется из Сигма X. Диапазон: 0-100.

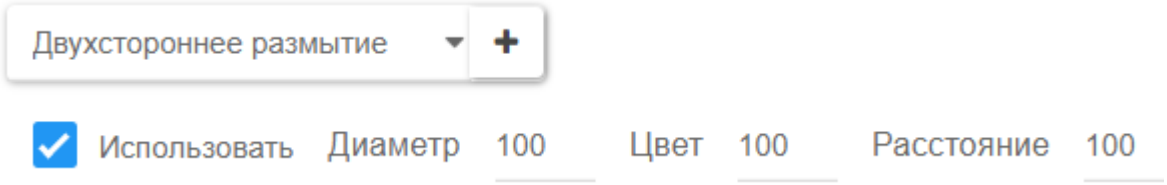
13) Двухстороннее размытие.

Данный этап предобработки выполняет двухсторонний фильтр с указанными настройками. При данном этапе происходит размытие фона изображения, при этом основные элементы не размываются. Этап работает только на цветных изображениях.

Пример использования:

(Рис. 35 Пример использования фильтра «Двухстороннее размытие»)

Интерфейс настройки.



(Рис. 36 Настройка фильтра «Двухстороннее размытие»)

Использовать. Если эта опция выбрана, то к изображению будет применен этап предобработки – Двухсторонний фильтр.

Диаметр. Диаметр каждой окрестности пикселя, который используется во время фильтрации. Диапазон: 1-100.

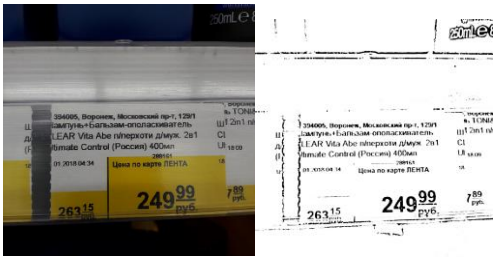
Цвет. Фильтр сигмы в цветовом пространстве. Большее значение параметра означает, что цвета в окрестности пикселя будут смешиваться вместе, в результате образуются большие области схожих цветов. Диапазон: 1-100.

Расстояние. Фильтр сигма в координатном пространстве. Большее значение параметра означает, что дальнейшие пиксели будут влиять друг на друга, если их цвета достаточно близки. Диапазон: 1-100.

14) Адаптивное устранение шумов

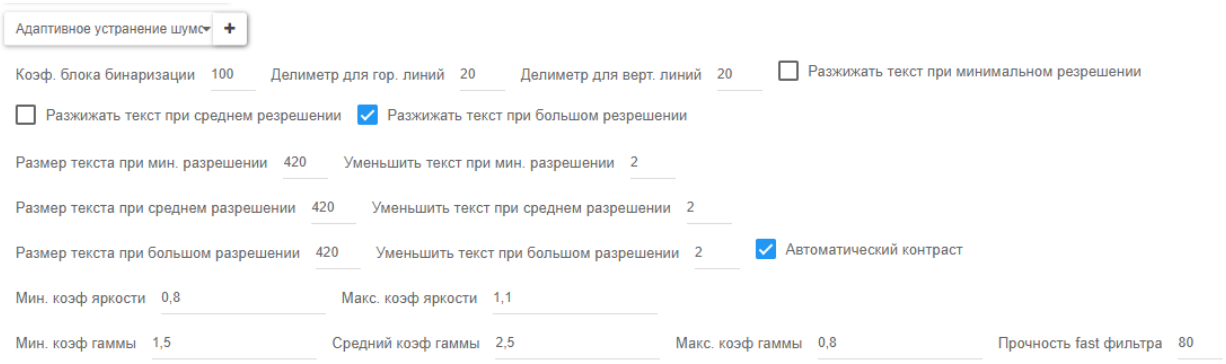
Данный этап предобработки сочетает изменение яркости изображения, удаление линий, бинаризацию изображения и морфологические преобразования с целью выделения текстовых областей независимо от фона.

Пример использования.



(Рис. 37 Пример использования фильтра «Адаптивное устранение шумов»)

Интерфейс настройки. Параметры указываются автоматически. Вносить изменения вручную нет необходимости.



(Рис. 38 Настройка фильтра «Адаптивное удаление шумов»)

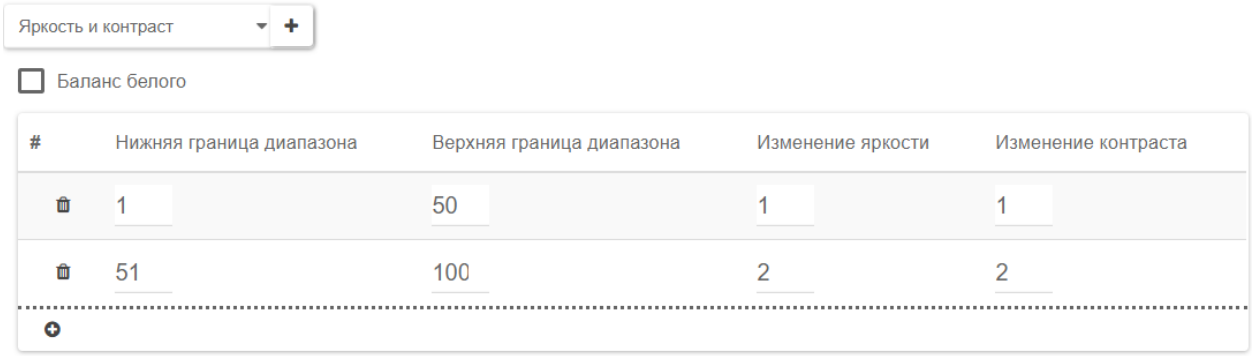
15) Яркость и контраст

Этап предобработки служит для адаптивного изменения яркости и контраста изображения. Так же в настройках можно применить фильтр «Баланс белого», который помогает выделить области белого фона и удалить шумы на изображении.

Пример использования:

(Рис. 39 Пример использования фильтра «Яркость и контраст»)

Интерфейс настройки.



(Рис. 40 Настройка фильтра «Яркость и контраст»)

Баланс белого. При включенной опции запускается процесс цветокоррекции, в результате которой объекты, которые глаз видит как белые, будут показаны белыми. Пограничные с ним объекты так же корректируют оттенок. Настройки коллекции диапазонов в этом случае не работают. Данную опцию можно использовать для удаления шумав.

Коллекция диапазонов яркостей. Указывает настройки изменения яркости и контраста изображения для тех изображений, рассчитанная яркость которых попадает в указанный диапазон.

Нижняя граница диапазона. Значение яркости изображения, ниже которого указанные настройки не будут применены. Диапазон: 0-100.

Верхняя граница диапазона. Значение яркости изображения, выше которого указанные настройки не будут применены. Диапазон: 0-100.

Диапазоны не могут пересекаться. Нижняя граница не может быть больше верхней границы.

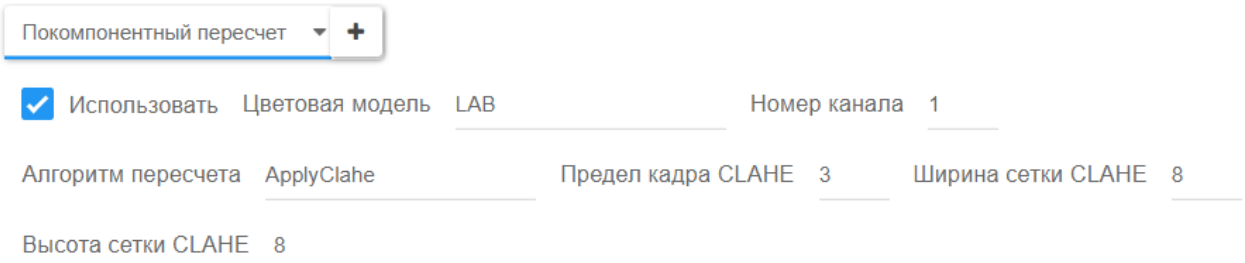
Изменение яркости. Указывает значение, на которое будет увеличена яркость изображения, при попадании текущей яркости в указанный диапазон. Диапазон: -255 – 255.

Изменение контраста. Указывает значение, на которое будет увеличен контраст изображения, при попадании текущей яркости в указанный диапазон. Диапазон: -127 – 512.

16) Покомпонентный пересчет

Данный этап предобработки позволяет применить один из алгоритмов нормализации к одному из каналов в указанной цветовой модели.

Интерфейс настройки.



(Рис. 41 Настройка фильтра «Покомпонентный пересчет»)

Использовать. Если эта опция выбрана, то к изображению будет применен этап предобработки – Покомпонентный пересчет.

Цветовая модель. Указывает цветовую модель, на которую будет раскладываться изображение для выбора канала. Варианты: HSV, YCrCb, LAB:

- **HSV**. Каналы содержат информацию о, соответственно: оттенке, насыщенности и яркости.
- **YCrCb**. Каналы содержат информацию о, соответственно: яркости, красной цветоразностной компоненте и синей цветоразностной компоненте.
- **LAB**. Каналы содержат информацию о, соответственно: «светлоте», положение цвета от зеленого до красного и положение цвета от синего до желтого.

Номер канала. Номер канала цветовой модели к которому будет применен нормализующий алгоритм. Диапазон: 1-3.

Алгоритм пересчета. Указывает алгоритм нормализации канала изображения. Варианты: EqualizeHist, ApplyClahe:

- **EqualizeHist**. Это метод улучшает контрастность изображения, растягивая диапазон интенсивности.
- **ApplyClahe**. Этот метод выравнивает контраст по ограниченной адаптивной гистограмме.

Предел кадра CLAHE. Порог для ограничения контрастности для алгоритма CLAHE. Диапазон: 0-100.

Ширина сетки CLAHE. Ширина сетки для выравнивания гистограммы. Диапазон: 0-100.

Высота сетки CLAHE. Высота сетки для выравнивания гистограммы. Диапазон: 0-100.

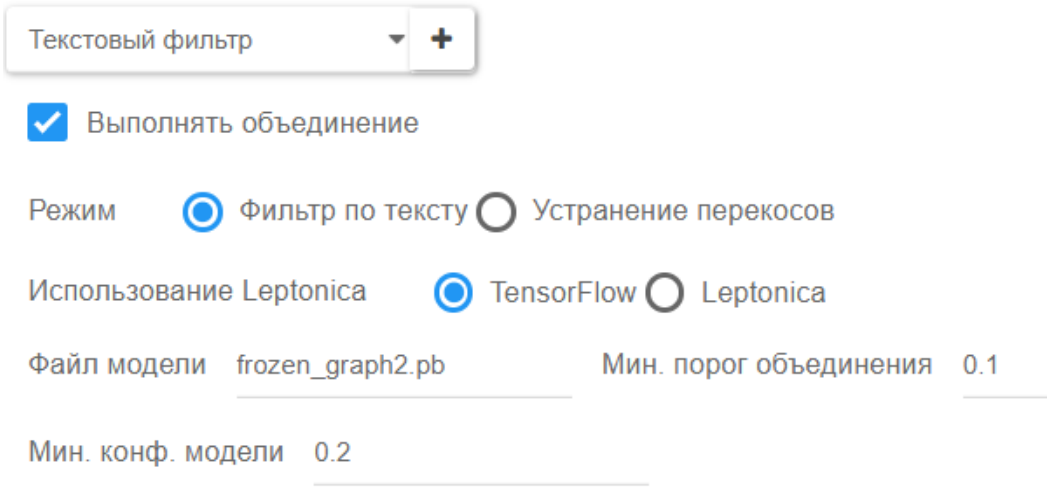
Входное изображение будет разделено на прямоугольные плитки одинакового размера.

17) Текстовый фильтр.

Данный этап предобработки выполняет поиск текстовых областей на репрезентации с помощью нейросети, и наложение полученной из найденных прямоугольников маски на изображение.

Нейросети используются по натренированным моделям TensorFlow. После тренировки нейросети файл с расширением .pb располагается на сервере и выбирается внутри локатора.

Интерфейс настройки



(Рис. 42 Настройка фильтра «Текстовый фильтр»)

Выполнять объединение. Опция, которая указывает на то, будут ли объединяться накладывающиеся текстовые блоки.

Имя файла модели нейронной сети. Имя файла, который должен находиться в корневой папке приложения с обученной моделью для нейросети.

Минимальный процент доверия нейросети. Параметр указывающий минимальную степень доверия к результату поиска текстового блока, с которой этот блок будет учитываться. Диапазон значений: 0-100.

Порог объединения. Указывает процент пересечения текстовых блоков для их объединения. Диапазон значений: 0-100.

Режим устранения перекосов. Опция, которая указывает на то, будет ли выполняться закрашивание областей изображения без текста или изображение будет повернуто на вычисленный из координат слов текста доминантный угол поворота (только для поиска текста при помощи нейросети).

Использовать Leptonika. Опция, указывающая будут ли текстовые области искаться при помощи нейросети или Leptonika.

Минимальная ширина слова Leptonika. Указывает минимальную ширину слова в пикселях при поиске текста с помощью Leptonika.

Минимальная высота слова Leptonika. Указывает минимальную высоту слова в пикселях при поиске текста с помощью Leptonika.

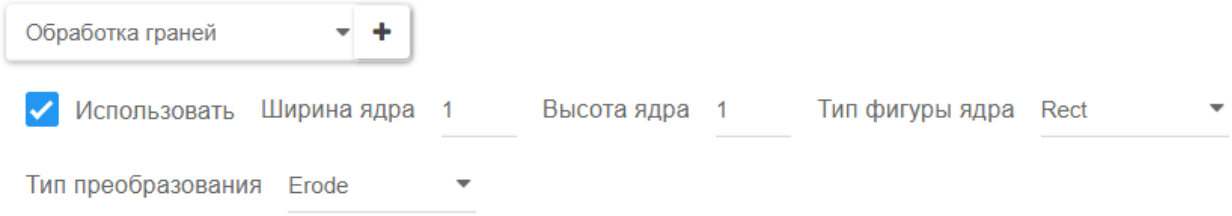
Максимальная ширина слова Leptonika. Указывает максимальную ширину слова в пикселях при поиске текста с помощью Leptonika.

Максимальная высота слова Leptonika. Указывает максимальную высоту слова в пикселях при поиске текста с помощью Leptonika.

18) Обработка граней.

Выполняет одно из указанных преобразований с определенной структурой указанного размера. Все объекты на изображении обрабатываются в зависимости от указанного типа преобразования. Их границы сглаживаются, либо ужирняются, либо выделяется только контур объекта.

Интерфейс настройки.



(Рис. 43 Настройка фильтра «Обработка граней»)

Использовать. Если эта опция выбрана, то к изображению будет применен этап предобработки – Морфологические преобразования.

Ширина ядра. Значение указывающая ширину блока для морфологических преобразований, выраженная в процентах от ширины изображения. Диапазон: 0,01-100.

Высота ядра. Значение указывающая высоту блока для морфологических преобразований, выраженная в процентах от высоты изображения. Диапазон: 0,01-100.

Тип фигуры ядра. Указывает форму «клише» (структуры) для морфологических преобразований. Варианты: прямоугольник, эллипс, перекрестие.

Тип преобразования. Указывает алгоритм, по которому выполняется преобразование изображения с помощью структуры указанной формы и размера. Варианты: Erode, Dilate, Open, Close, Gradient, Top Hat, Black Hat, Hit Miss:

Erode. В данном случае толщина линий объектов увеличивается, т.е. символы «обрастают» пикселями.

Dilate. В данном случае фон размывает, «выедает» границы объектов.

Open. Здесь выполняется Dilate результата Erode.

Close. Здесь выполняется Erode результата Dilate.

Gradient. Здесь выполняется вычитание из Dilate результата Erode. Т.е. нахождение границ.

Top Hat. Здесь из исходного изображения вычитаем результат Open.

Black Hat. Здесь из результата Close вычитаем исходное изображение.

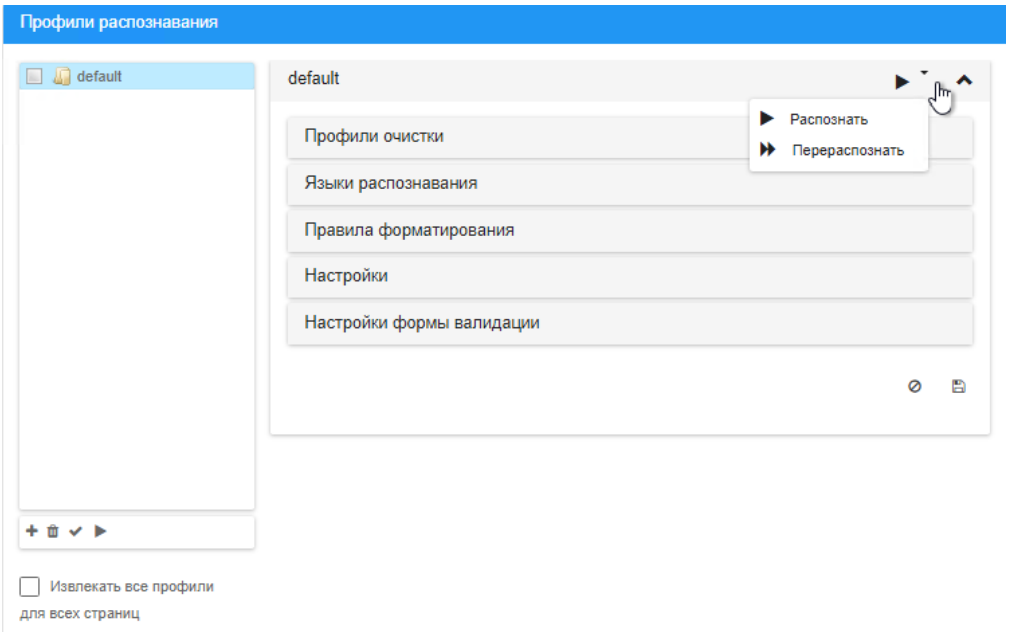
Hit Miss. В данной версии не используется.

2.3.2. Профили распознавания.

Профиль распознавания – набор конфигурационных настроек, которые будут применяться при распознавании изображения выбранным профилем. Необходимо выбрать алгоритм «Распознать» или «Перераспознать» с помощью стрелок.

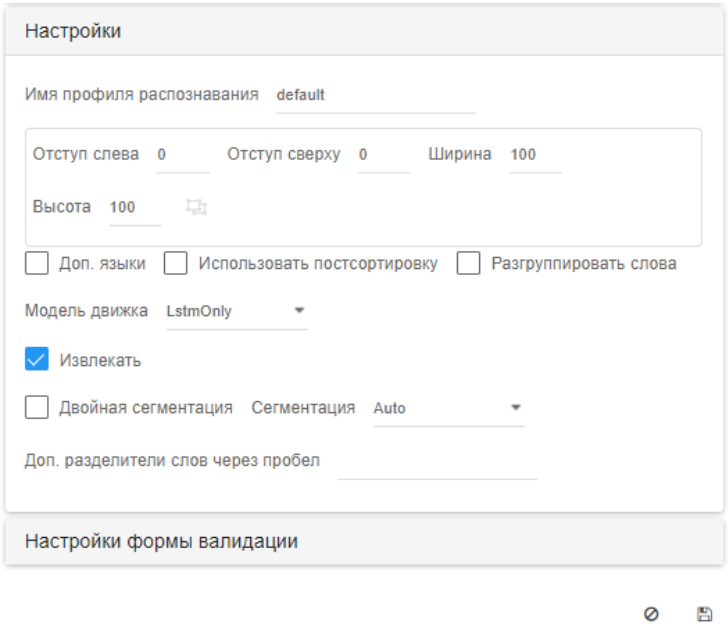
Результаты работы разных профилей могут сравниваться между собой для выявления лучшего итогового результата – что повышает качество распознавания.

Форма профилей распознавания делится на 2 части:



(Рис. 44 Меню профилей распознавания)

— **Общий список профилей.** С помощью меню общего списка (+ 🗑 ✓ ▶) можно Создать новый профиль, Удалить выбранный профиль, Выделить несколько профилей, Запустить выделенные профили.
— **Настройка выбранного/нового профиля.**



(Рис. 45 Настройка профиля распознавания)

Имя профиля – задается/изменяется имя.

Выбор региона распознавания:

Отступ слева - отступ от левой границы изображения репрезентации выраженный в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.

Отступ сверху - отступ от верхней границы изображения репрезентации выраженный в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.

Ширина – ширина области, выраженная в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.

Высота – высота области, выраженная в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.

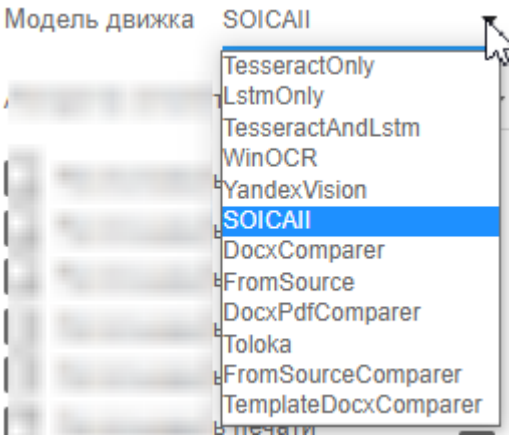
Суммы Отступ слева, Ширина и Отступ сверху, Высота так же не могут быть больше 100.

Дополнительные языка. Если эта опция выбрана, то помимо основных языков для получения OCR будут использованы дополнительные, имя которых должно начинаться с имени основного языка, например, «eng2» или «rus_kursiv». Могут применяться только в движках Tesseract и LSTM.

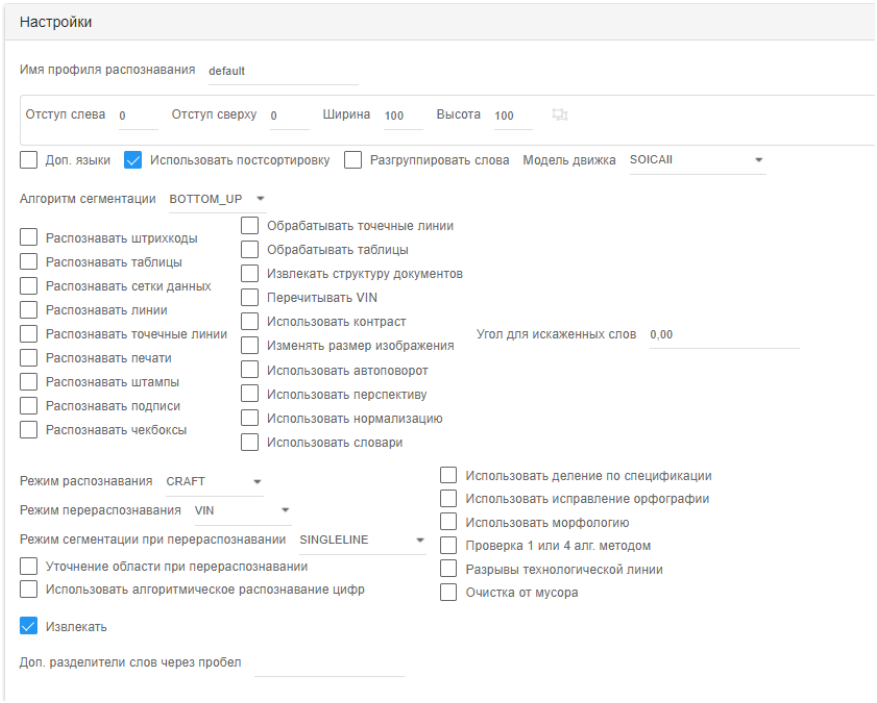
Использовать постсортировку. При выборе этой опции после получения OCR происходит сортировка слов по местоположению.

Разгруппировать слова. При использовании этой опции исходная информация о текстовых линиях стирается и линии заново собираются (все слова состоят из набора линий). Если опция не выбрана текстовые линии не собираются.

Модель движка. Указывает модель, по которой будет производиться получение OCR. Варианты: TesseractOnly, LstmOnly, TesseractAndLstm, WinOCR, YandexVision, SOICAII, DocxComparer, FromSource, DocxPdfComperer, Toloka, FromSourceComparer, TemplateDocxComparer.



- **TesseractOnly**. Используется только Tesseract.
- **LstmOnly**. Используется только LSTM.
- **TesseractAndLstm**. Выбор движка осуществляется автоматически из Tesseract и LSTM.
- **WinOCR.** При использовании необходимо выбирать русский язык, методы сегментации не влияют на результат распознавания.
- **SOICAII.** При использовании необходимо выбирать язык и необходимые опции.
- **DocxComparer.** Позволяет сравнить 2 документа в формате .docx и получить документ-результат сравнения в формате .docx, в котором будут выделены удаленные и добавленные слова и фразы.
- **FromSource.** Позволяет извлекать OCR из импортированного файла. В настройках необходимо выбрать язык. Если файл-источник имеет формат .pdf с текстом или .docx-документ, то для этого профиля сразу на импорте будет создана репрезентация с текстом из источника (распознавание производится не будет). Также, если этот движок используется для экспорта пакета (целиком, не по документам) в pdf, то в качестве экспортируемого pdf будет выгружаться исходный файл.
- **DocxPdfComparer.** Позволяет сравнить 2 документа (версии одного документа), один из которых формата .docx, а второй имеет формат .pdf и получить документ-результат сравнения в формате .docx, в котором будут выделены удаленные и добавленные слова и фразы.
- **Toloca.** Распознавание на основе сервиса Toloca.
- **FromSourceComparer.** Позволяет сравнить 2 документа (версии одного документа) формата .docx, .xlsx, .pdf и получить документ-результат сравнения в формате .docx, в котором будут выделены удаленные и добавленные слова и фразы.
- **TamplateDocxComparer.** Производит сравнение шаблона с заполненным документом в формате .docx и получить документ-результат сравнения в формате .docx. Шаблоном считается документ, в названии которого присутствует слово tamplate.



(Рис. 45.1 Настройки профиля распознавания с движком SOICAII)

Извлекать. Данная опция указывает, может ли быть выполнено получение репрезентации данным профилем для страницы изображения и, соответственно, могут ли результаты использоваться в инструментах извлечения, классификации, экспорта и прочих.

Двойная сегментация. Если выбрана данная опция, то OCR будут получено два раза: для получения линий (указанным методом сегментации) и текста (методом сегментации Auto). Может применяться только в движках Tesseract и LSTM.

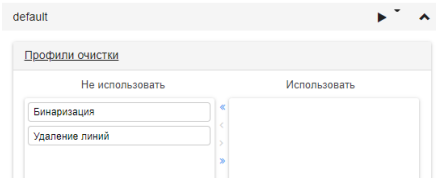
Сегментация. Указывает метод формирования текстовых блоков при получении OCR. Может применяться только в движках Tesseract и LSTM.

Варианты: OsdOnly, AutoOsd, AutoOnly, Auto, SingleColumn, SingleBlockVertText, SingleBlock, SingleLine, SingleWord, CircleWord, SingleChar, SparseText, SparseTextOsd, RawLine:

- **OsdOnly**. Ориентация и обнаружение скриптов (OSD).
- **AutoOsd**. Автоматическая сегментация страницы с OSD.
- **AutoOnly**. Автоматическая сегментация страниц, но не OSD, или OCR.
- **Auto**. Полностью автоматическая сегментация страницы, но без OSD. (По умолчанию)
- **SingleColumn**. Предполагается, что один столбец текста переменных размеров.
- **SingleBlockVertText**. Предполагается, что единый блок вертикально выровненного текста.
- **SingleBlock**. Предполагается, что единый блок текста.
- **SingleLine**. Рассматривать изображение в виде одной строки текста.
- **SingleWord**. Рассматривать изображение как одно слово.
- **CircleWord**. Рассматривать изображение как одно слово в круге.
- **SingleChar**. Рассматривать изображение в виде одного символа.
- **SparseText**. Разреженный текст. Найти как можно больше текста в определенном порядке.
- **SparseTextOsd**. Разреженный текст с OSD.

- **RawLine**. Сырая линия. Обработать изображение как одну текстовую строку, минуя разделители, специфичные для Tesseract.

Дополнительные разделители слов через пробел. Указывает список совокупностей символов, по которым выполняется разделение слов из результатов OCR.

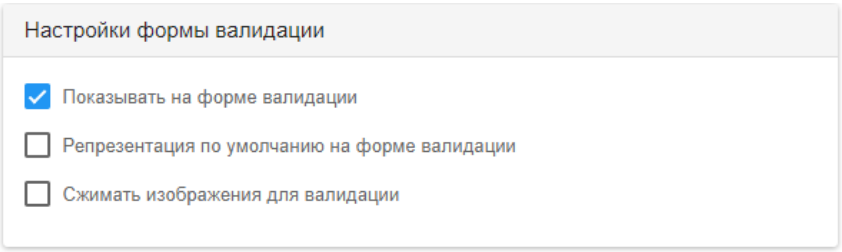


Для каждого профиля распознавания может быть применено несколько профилей очистки. Они выбираются из списка имеющихся профилей очистки выбранного класса пакета.

Языки распознавания выбираются аналогичным способом. Указанными языками будет выполняться получение OCR. Список может быть пустым. Первый элемент списка является главным, т.е. будут использованы настройки движка распознавания, указанные для этого языка.

Так же может применятся правила форматирования. Список правил форматирования последовательно применяется к словам результатов OCR репрезентации.

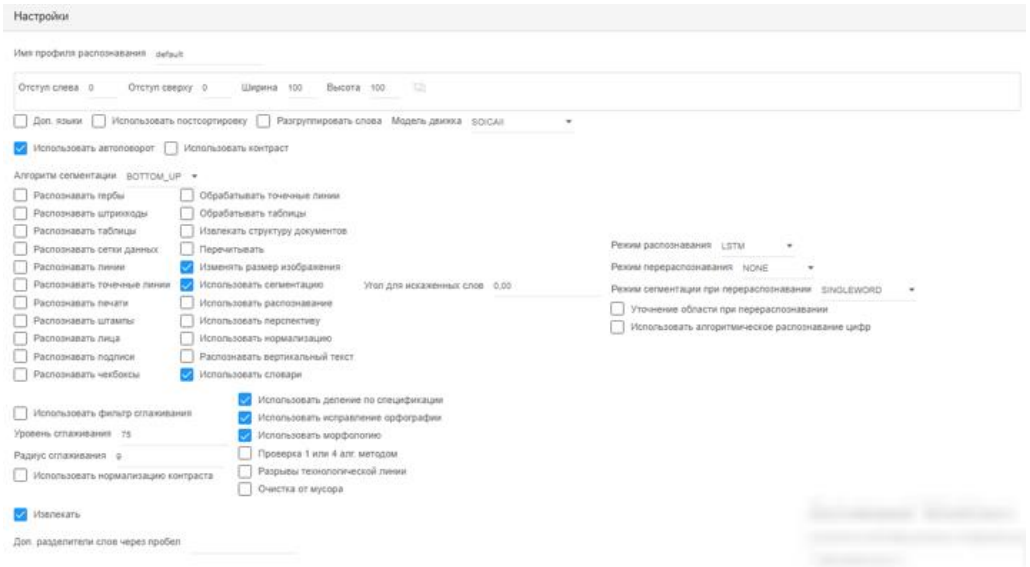
В настройках формы валидации можно выбрать отображение данного профиля распознавания на форме валидации.



Варианты: Показывать на форме валидации, Репрезентация по умолчанию на форме валидации; Сжимать изображения для Валидации.

Настройки движка SOICAII

Настройка движка SOICAII гибкая, простая и интуитивно понятная.

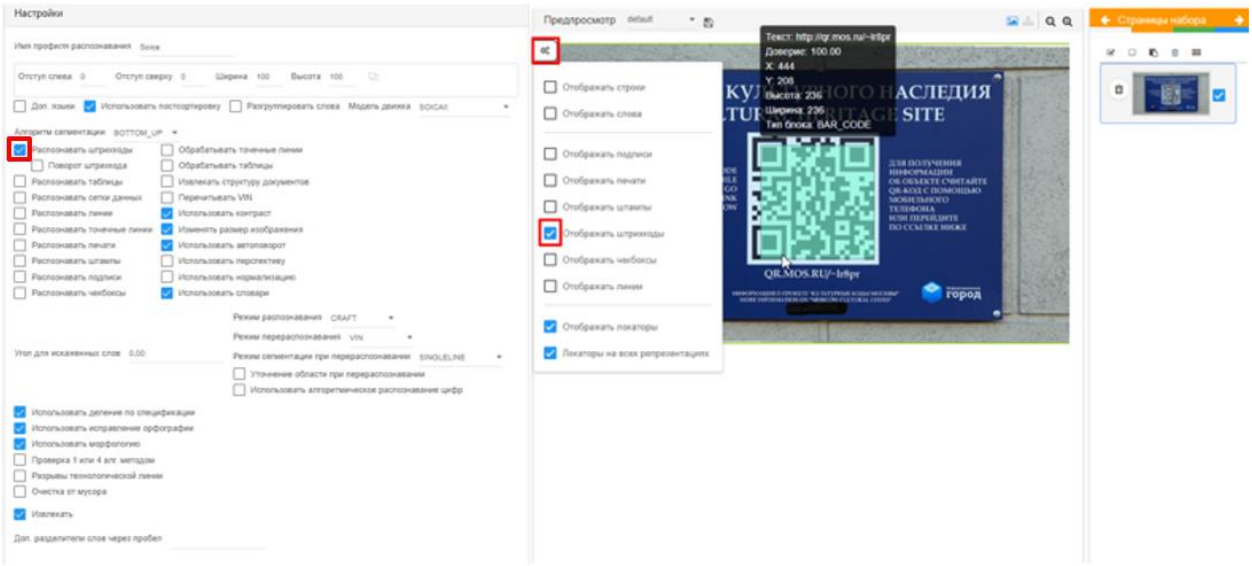


(Рис. 45.2 Настройка профиля распознавания)

В настройках профиля распознавания на базе движка SOICAII есть возможность опционально выбрать дополнительные элементы распознавания. Результаты части дополнительных элементов распознавания движка SOICAII можно увидеть в окне предпросмотра, нажав на шестеренки и выбрав нужную опцию.

Дополнительные элементы распознавания движка SOICAII:

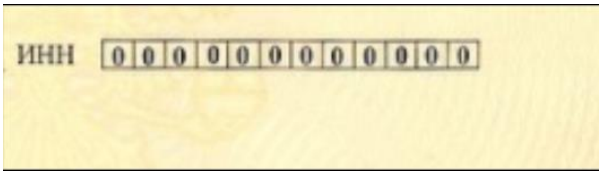
- **Распознавать штрих-коды;**



(Рис. 45.3 Настройка и отображение поиска штрих кодов)

- **Распознавать таблицы**;

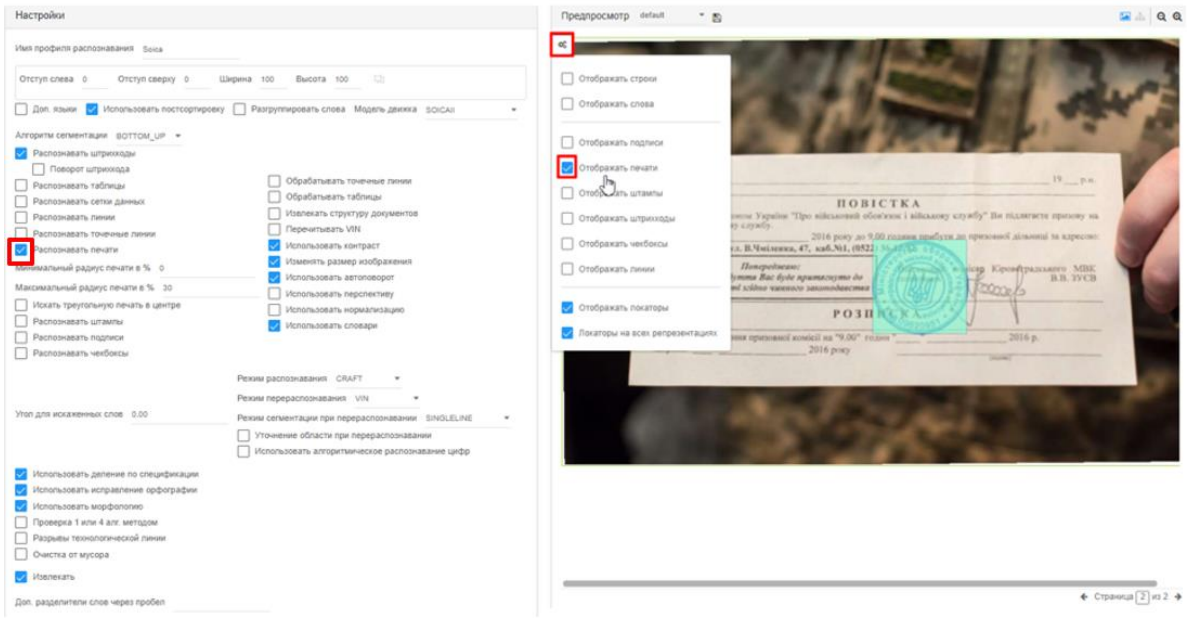
- **Распознавать сетки данных.** Это какие-либо данные занесенные в сетку, т е упорядоченные и оформленные в виде строки разделенной линиями. Пример:



(Рис. 45.4 Пример сетки данных)

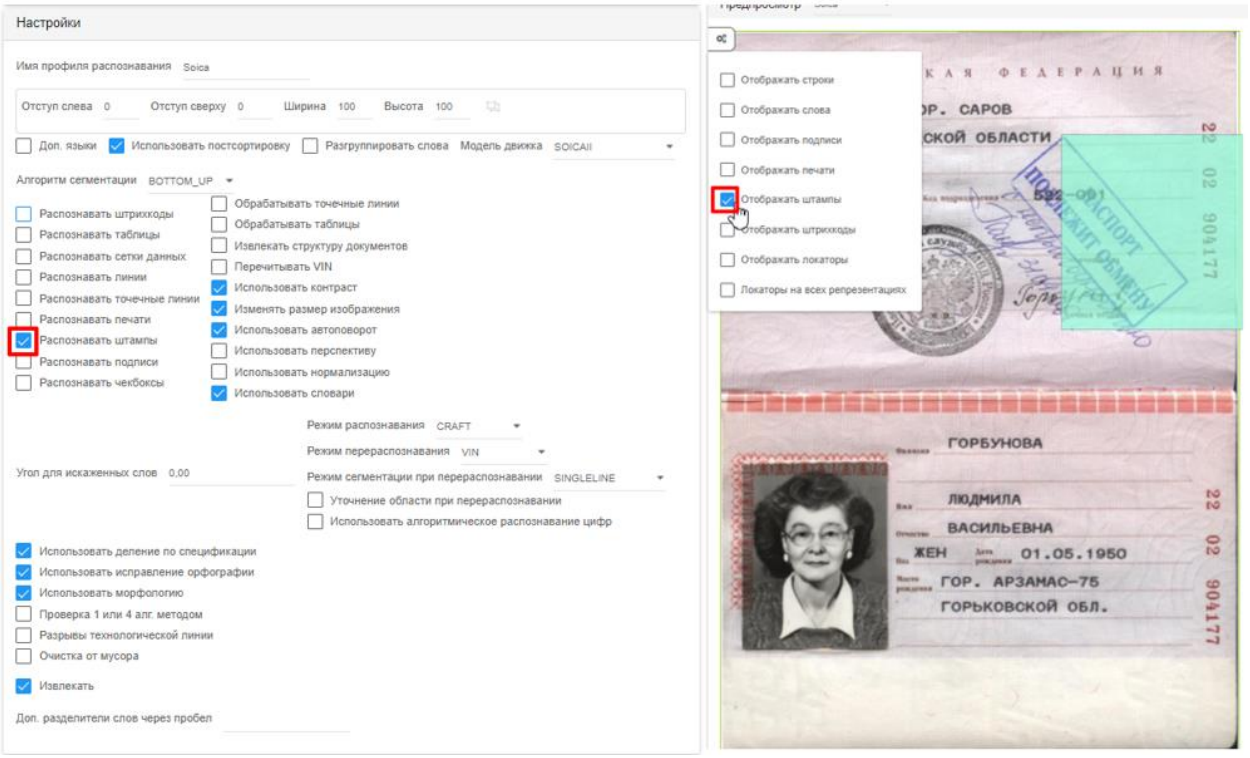
- **Распознавать линии**;

- **Распознавать печати**;



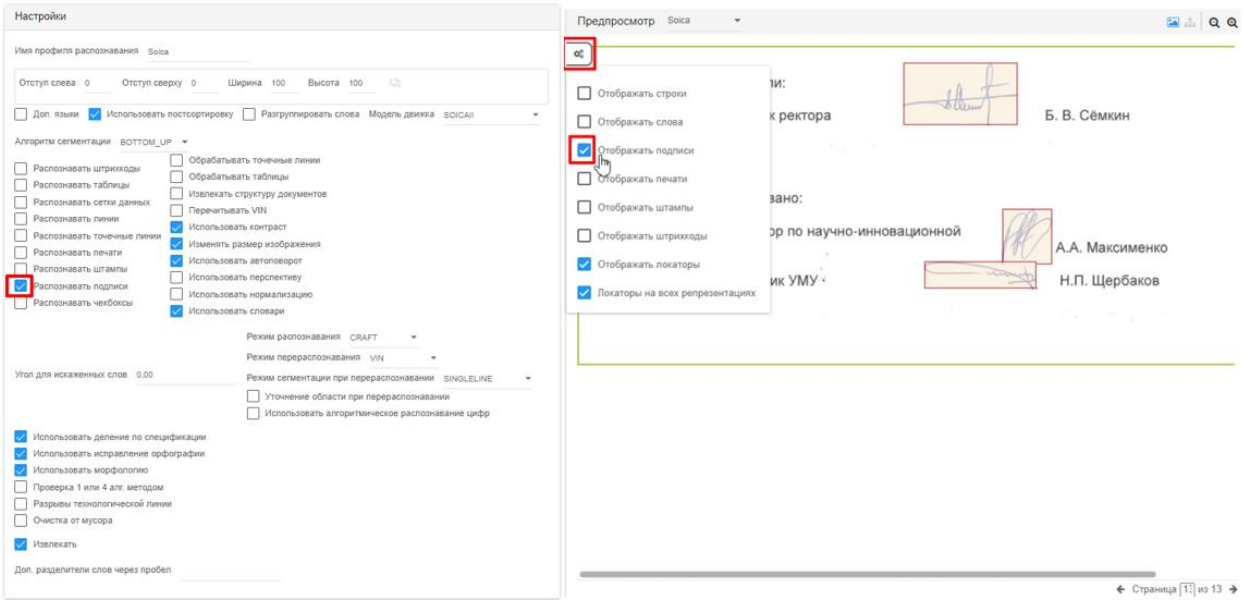
(Рис.45.5 Настройка и отображение поиска печатей)

- **Распознавать штампы**;



(Рис.45.6 Настройка и отображение поиска штампов)

- **Распознавать подписи**;



(Рис. 45.7 Настройка и отображение поиска подписей)

- **Распознавать чекбоксы**. Производится поиск чекбоксов.

- **Распознавать гербы.** Осуществляется поиск гербовых орлов.

Также, в зависимости от вида искомых данных и качества изображения, можно выбрать дополнительные настройки обработки изображения:

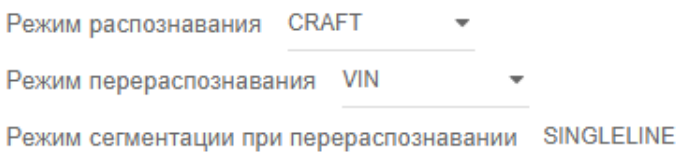
- **Использовать контраст.** Происходит автоматическое улучшение изображения.

- **Использовать автоповорот.** Выполняет автоповорот изображения на угол кратный 90 градусов по тексту. Поворот производится относительно найденного текста.

- **Использовать перспективу.** Производится устранение сложных перекосов с нарушением геометрии изображения.

- **Использовать словари.** Включение постобработки результатов OCR движка SOICAII, в том числе исправление незначительных недостатков OCR с помощью словарей;

- **Перечитывать VIN**. При выборе этой опции происходит пере распознавание определенного фрагмента с помощью специально обученных моделей (нейросетей). Дополнительно необходимо произвести настройки блока:

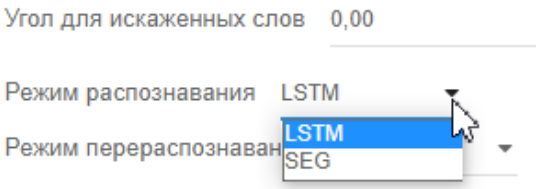


(Рис. 45.8 Настройки перечитывания фрагмента)

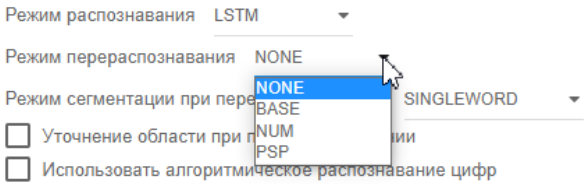
- **Распознавать точечные линии**. Поиск точечных линий и вывод их в выходных данных

- **Извлекать структура документа.** При выборе данной опции извлекается вся структура документа (иерархия пунктов и подпунктов в документе), данная структура может извлекать данные сразу с нескольких страниц).

Режим распознавания – модели движков для пере распознавания. Варианты: LSTM, SEG.

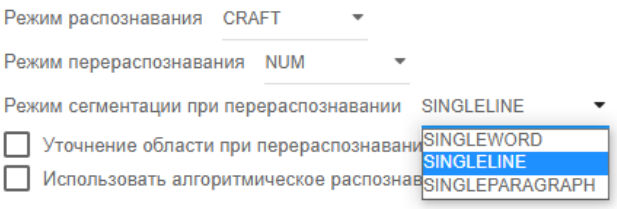


Режим перераспознавания – алгоритм пере распознавания. Варианты: Base (пере распознавание стандартной модели LSTM), NUM, PSP.



Для режима распознавания SEG необходимо выбирать режим перераспознавания PSP или NUM. Оба этих режима подойдут для перераспознавания серии и номера паспорта, отличаются по шрифтам.

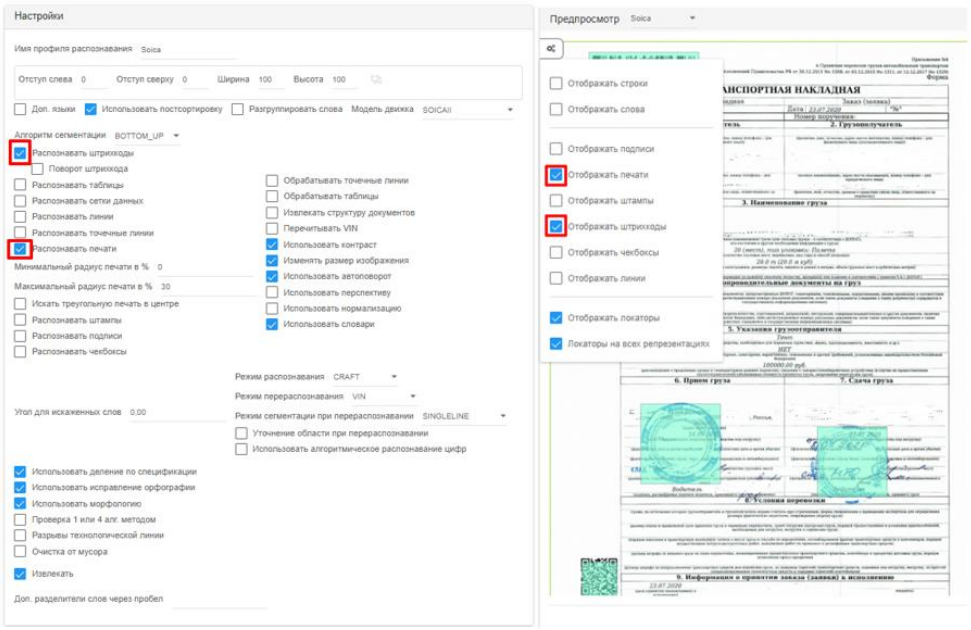
Режим сегментации при перераспознавании – типы сегментации при пере распознавании фрагмента изображения.



Опция **Угол для искаженных слов** используется, например, при распознавании чертежей, в которых текст расположен под углом к линиям.

Опция **Использовать фильтр сглаживания** выполняет предобработку изображения с использованием адаптивной нормализации.

Одновременно можно выбрать несколько опций.



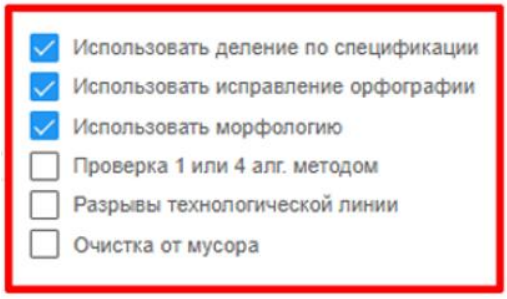
(Рис. 45.9 Настройка и отображение нескольких опций)

В случае, если ни одна из опций из списка не выбрана, будет происходить получение OCR в зависимости от выбранного языка.

Уточнение области при перераспознавании. При выборе опции происходит автоматическое уточнение (подгонка к оптимальной) области перераспознавания.

Использовать алгоритмическое распознавание цифр. Выполняется алгоритмическое распознавание цифр.

В настройках профиля распознавания Soica есть отдельный блок настроек, который применяется к уже полученному OCR.



(Рис. 45.10 Настройки, применяемые к уже полученному OCR)

Использовать деление по спецификации. Используется деление по спецсимволам.

Использовать исправление орфографии. Исправление опечаток или исправление результатов OCR на основании использования словарей.

Использовать морфологию. Позволяет использовать морфологические конструкции для поиска данных.

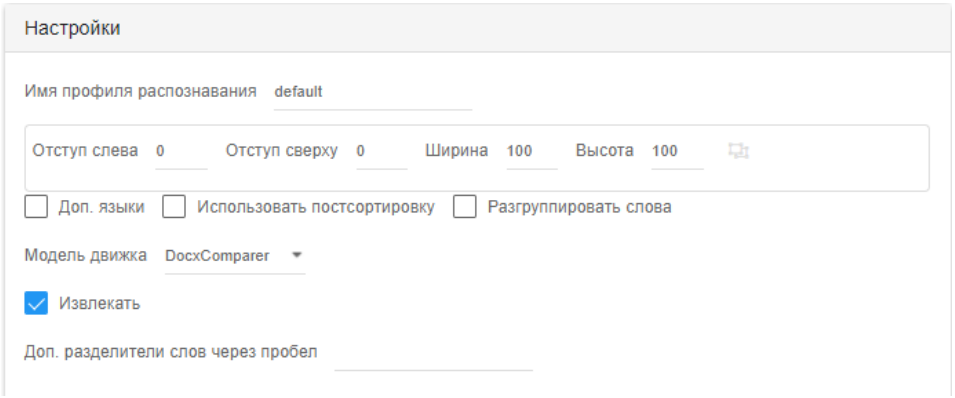
Проверка 1 или 4 алг. методом. Выполняется проверка OCR на корректность полученных цифр 1 и/или 4.

Разрывы технологической линии. Происходит объединение слов с переносами.

Очистка от мусора. По результатам OCR убирается явный не корректный мусор. Например, большие фрагменты печати распознались как набор хаотичных спецсимволов.

Сравнение документов

Для сравнения документов в основных настройках профиля распознавания необходимо выбрать модель движка DocComparer. Движок DocComparer позволяет сравнить 2 документа в формате docx и получить документ-результат сравнения в формате docx, в котором будут выделены удаленные и добавленные слова и фразы.



(Рис. 45.10 Настройка сравнения документов)

Работа с 2-мя репрезентациями

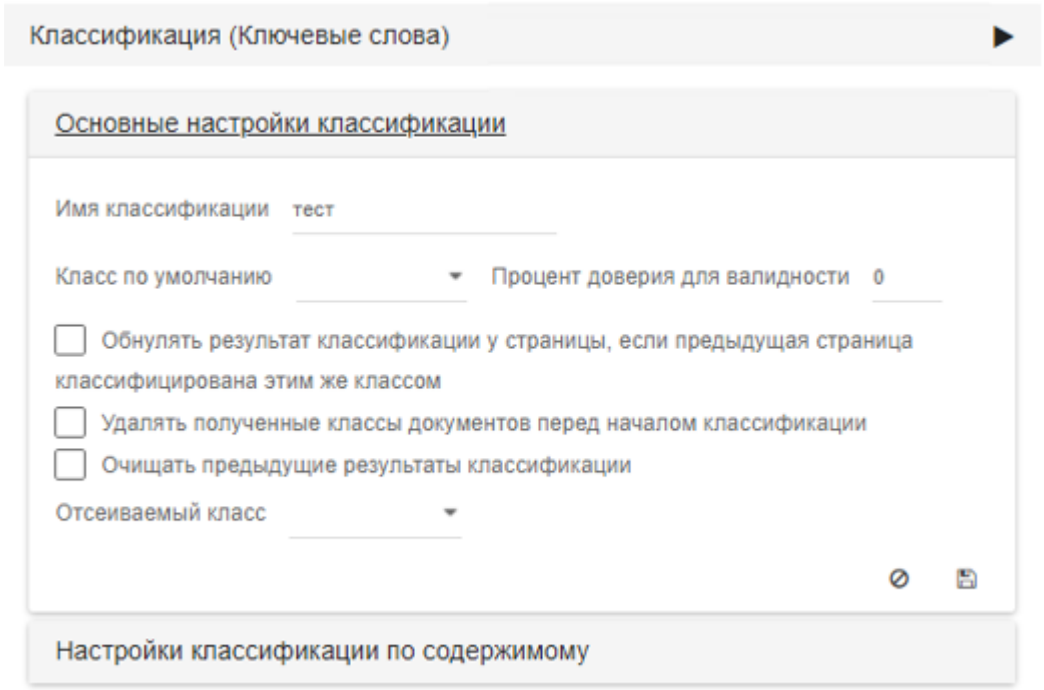
Движок SOICAII возвращает 2 репрезентации. К координатам на 1-ой репрезентации привязан текст и таблицы, а к координатам на 2-ой репрезентации - остальные блоки SOICAII. Обе репрезентации привязаны к одному профилю распознавания, поэтому в локаторах по-прежнему просто указывается профиль, с которого надо получать данные. Система автоматически берет нужные данные с нужной репрезентации.

2.3.3. Классификация.

Классификация документа определяет тип страниц (счет фактура, счет на оплату, паспорт и т.д.) и логику их дальнейшей обработки. А также объединяет многостраничные документы в один.

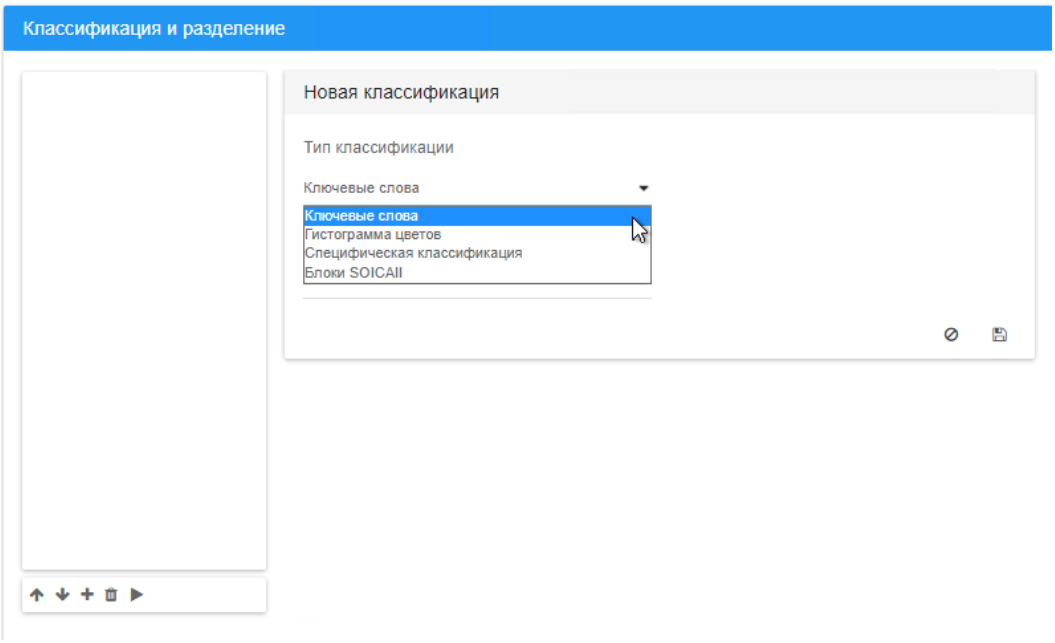
Классификация запускается нажатием кнопки 

Основные настройки классификации.



(Рис. 46 Основные настройки классификации)

Тип классификации. Указывает метод для получения кандидатов в результаты классификации. Выделяется 4 метода классификации: Ключевые слова, Гистограмма цветов, Специфическая классификация, Блоки SOICA. Методы классификации можно использовать отдельно, комбинировать или объединять между собой.

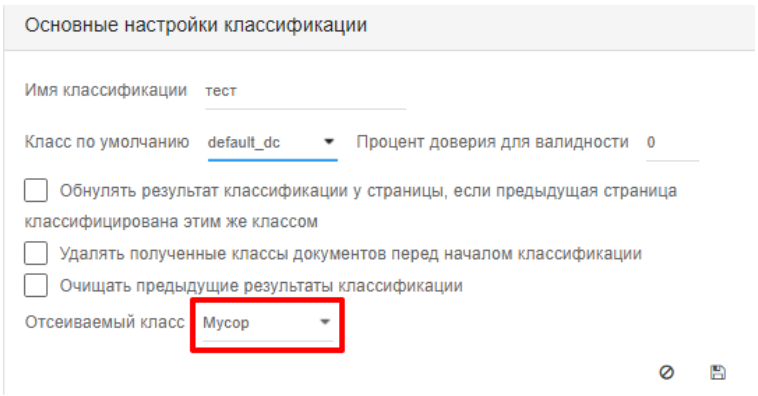


Класс по умолчанию. Указывает класс документа, который присваивается странице если ей не присваивается класс из кандидатов.

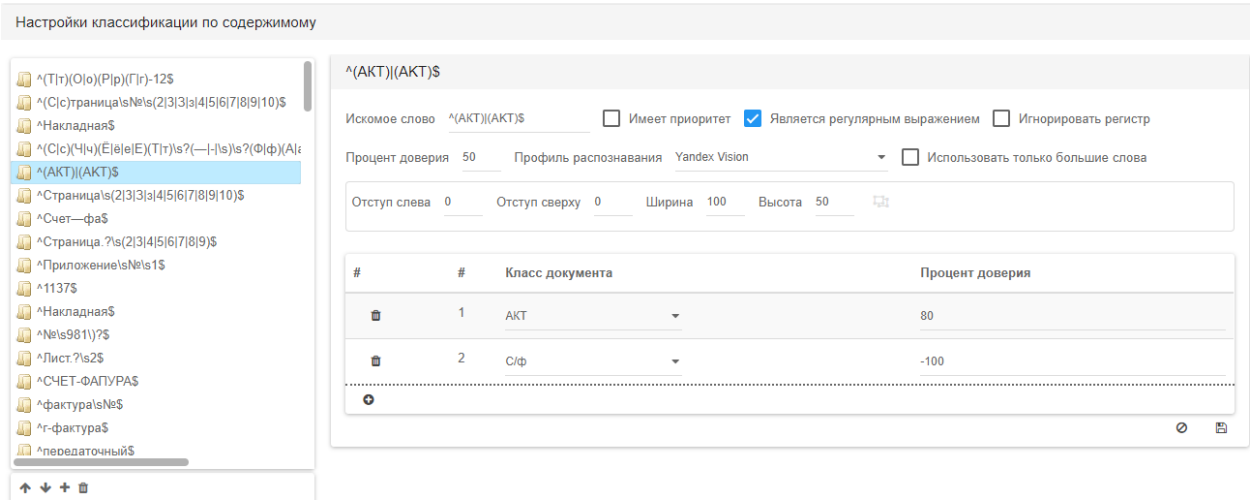
Процент доверия для валидности. Указывает минимальную степень доверия наилучшего (по степени доверия) кандидата в результаты классификации, при которой класс этого кандидата в результаты присваивается странице пакета. Диапазон значений: 0-100.

С помощью выбора соответствующей опции можно **Обнулять результат классификации у страницы, если предыдущая страница классифицирована этим же классом.**

Отсеиваемый класс. Позволяет перемещать в конец пакета страницы, классифицированные указанным классом. Пример:



Классификация по ключевым словам



(Рис. 47 Классификация, по ключевым словам)

В данном случае выполняется поиск соответствия указанных ключей со словами из результатов OCR в указанной области указанной репрезентации. Кандидаты в результаты классификации рассчитываются исходя из степени влияния ключа на результат по конкретному классу.

То есть каждый найденный ключ на документе определяет классифицируемую страницу к заданному типу на определённой количество процентов. Наибольшая итоговая сумма значения всех ключей присваивает странице класс.

Также можно классифицировать страницу импортируемого файла по её порядковому номеру. То есть первая страница любого импортируемого файла будет определяться как определенный класс документа, а остальные страницы будут проходить обычную классификацию по ключевому слову.

Пример 11:

Мы знаем, что на импорт будет поступать файл, в котором первая страница это список документов.

В настройках классификации мы указываем, что первая страница является документом класса «Список». Остальные страницы классифицируются по ключевым словам.

В итоге у любого файла, который обрабатывается этим сценарием первая страница будет классифицироваться как документ «Список».

Искомое слово. Указывает строку, которая будет сравниваться со словами из OCR.

Имеет приоритет. Если стоит данная отметка, то указанный ключ будет в приоритете при перед другими ключами. Его процент доверия к классу будет исключать остальные ключи.

Является регулярным выражение. Если стоит данная отметка, то искомое слово будет расцениваться как регулярное выражение (следовательно, оно должно быть записано в соответствующей форме).

Игнорировать регистр. Если стоит данная отметка, то не будет иметь значения заглавные или строчные буквы в искомом слове.

Процент доверия. Указывает минимальную степень доверия слова ключу при которой слово участвует в формировании кандидатов в результаты классификации. Процент рассчитывается как произведение степени доверия к слову в OCR и степени совпадения слова и ключа. Диапазон: 0-100.

Профиль распознавания. Указывает репрезентацию (результаты профиля распознавания) в которых будет осуществлен поиск ключей.

Использовать только большие слова. Если стоит данная отметка, то будут браться слова большие на 15% чем среднее по высоте символа.

Настройки области. Описывает область репрезентации, в которой будет выполняться поиск ключей.

- **Отступ слева** – отступ от левой границы изображения репрезентации выраженный в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.
- **Отступ справа** – отступ от верхней границы изображения репрезентации выраженный в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.
- **Ширина** – ширина области, выраженная в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.
- **Высота** – высота области, выраженная в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.

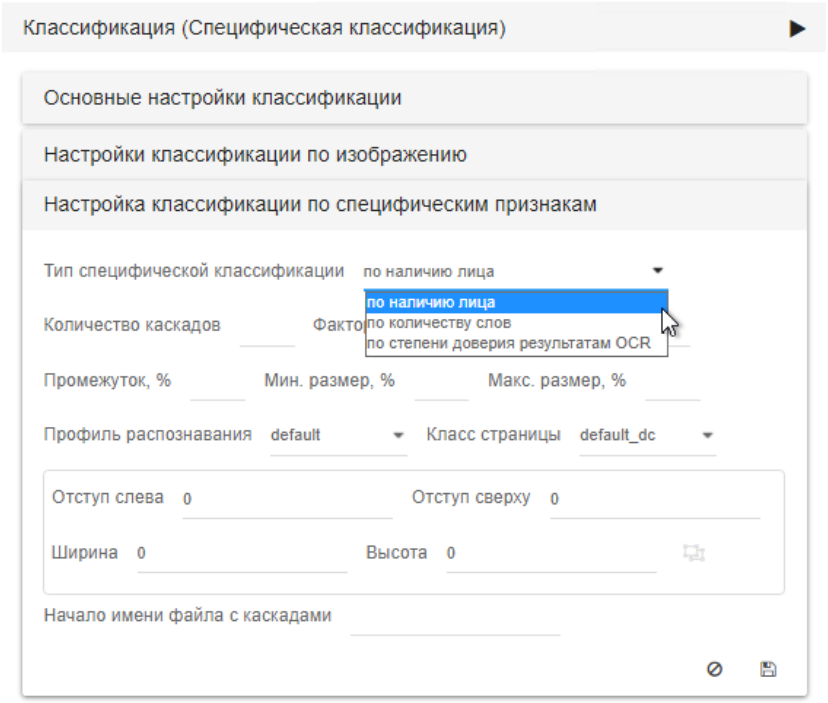
Суммы Отступ слева, Ширина и Отступ сверху, Высота так же не могут быть больше 100.

Класс документа. Указывает класс документа, значение кандидата которого будет изменено найденным ключом.

Процент доверия. Указывает процент доверия, участвующий в расчёте изменения степени доверия кандидата класса. Результат изменения рассчитывается как произведение доверия слова-ключа на степень влияния. Диапазон: -100 – 100.

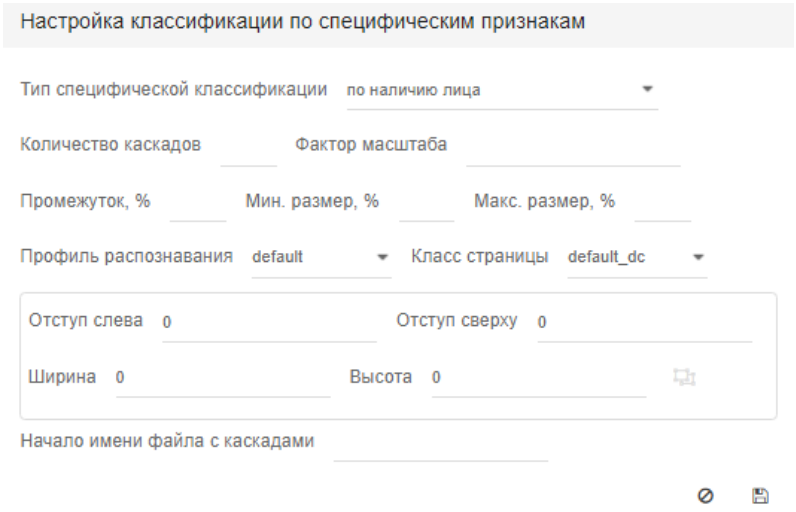
Специфическая классификация.

Содержит 4 подтипа: по каскадам Хаара, по наличию лица, по количеству слов, по степени доверия результатам OCR.

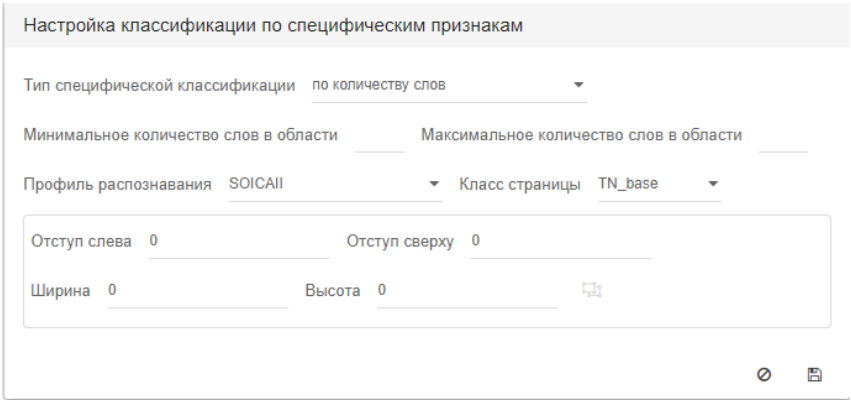


(Рис. 50 Типы специфической классификации)

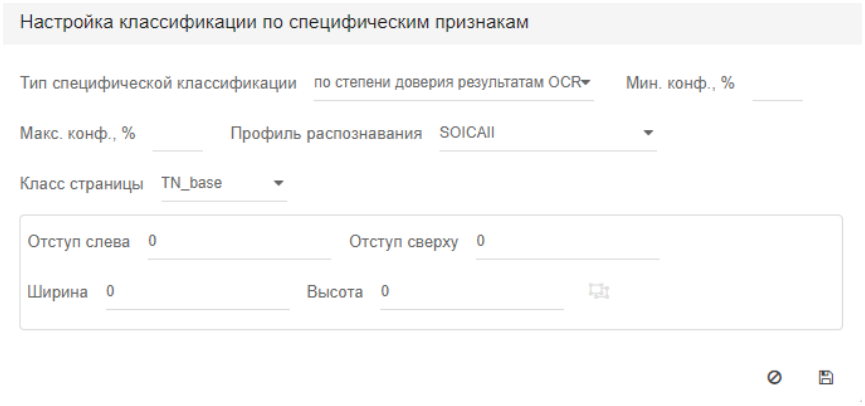
По наличию лица. В данном случае происходит поиск объекта по каскадам Хаара в указанной области указанной репрезентации. В случае нахождения объекта по одному из каскадов, степень доверия кандидата указанного класса увеличивается на 100%(количество используемых файлов с каскадами).



По количеству слов. Поиск по количеству слов в регионе.



По степени доверия результатам OCR. Поиск по среднему конфиденсу слов в регионе.



Профиль распознавания. Указывает репрезентацию в которых будет осуществлен поиск объектов по указанным каскадам.

Настройки области. Описывает область репрезентации, в которой будет выполняться поиск объектов по указанным каскадам.

- **Отступ слева** – отступ от левой границы изображения репрезентации выраженный в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.
- **Отступ сверху** – отступ от верхней границы изображения репрезентации выраженный в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.
- **Ширина** – ширина области, выраженная в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.
- **Высота** – высота области, выраженная в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.

Суммы Отступ слева, Ширина и Отступ сверху, Высота так же не могут быть больше 100.

Количество каскадов. Максимальное количество файлов с каскадами, по которым будет производится поиск объектов на репрезентации. Диапазон значений: 1-4.

Фактор масштаба. Параметр, показывающий на сколько будет меняться масштаб изображения, каждый проход поиска. Чем он меньше, тем дольше и подробнее будет выполняться поиск. Диапазон значений: 0,5-3.

Промежуток. Параметр, определяющий, сколько соседей должен иметь каждый прямоугольник-кандидат. Этот параметр влияет на качество обнаруженных лиц. Более высокое значение приводит к меньшему количеству обнаружений, но с более высоким качеством.

Минимальный размер. Указывает минимально возможный размер обнаружаемого объекта. Параметр выражается в процентах от ширины и высоты изображения. Диапазон значений: 1-100.

Максимальный размер. Указывает максимально возможный размер обнаружаемого объекта. Параметр выражается в процентах от ширины и высоты изображения. Диапазон значений: 1-100.

Начало имени файла с каскадами. Указывает начало наименования файлов с каскадами, которые будут использоваться для поиска объектов.

Классификация по цвету.

Выполняется расчет количества суммы пикселей по оттенкам в цветовой модели HSV, с кратностью в 10 градусов свернутой области изображения указанной репрезентации. Полученный результат сравнивается с настроенными моделями, и полученная степень совпадения гаммы участвует в расчете степени доверия кандидата классификации указанного класса.

Для выбора цветовой палитры необходимо выбрать документы в менеджере наборов и нажать кнопку «Сформировать», система автоматически определит основные цвета документа.

Данная классификация часто используется при работе с паспортами и иными документами имеющие характерные цвета.

Параметр «Влияние» - это параметр, показывающий как сильно влияет совпадение с указанной гистограммой на класс страницы.

Пример:

Класс А. Влияние = 50%. Гистограмма совпадает на 50%.

Класс Б. Влияние = 30%. Гистограмма совпадает на 90%.

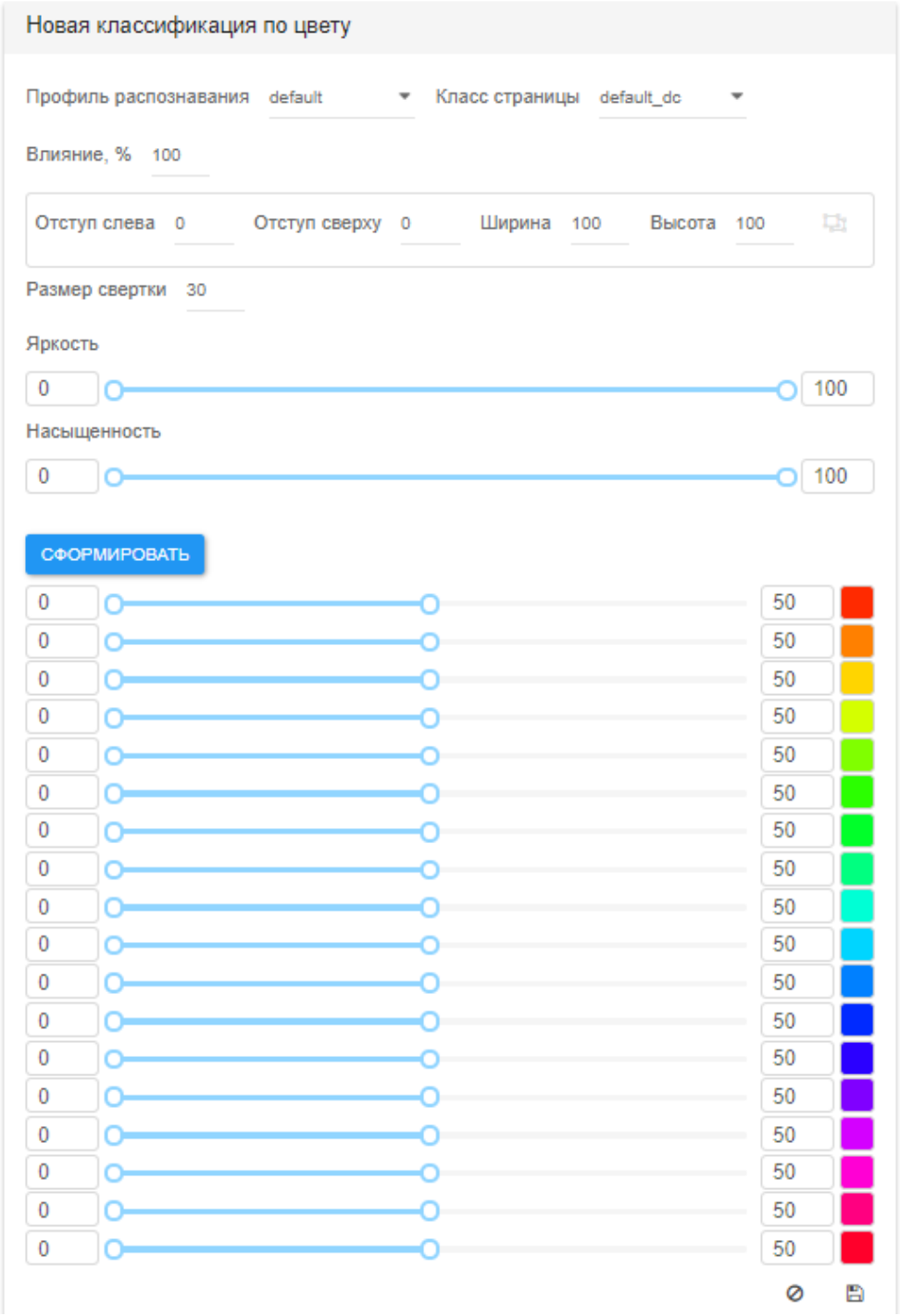
Класс С. Влияние = 90%. Гистограмма совпадает на 10%.

Кандидаты классов для страницы тогда:

А: 50%*50%=25% В: 30%*90%=27% С: 90%*10%=9%

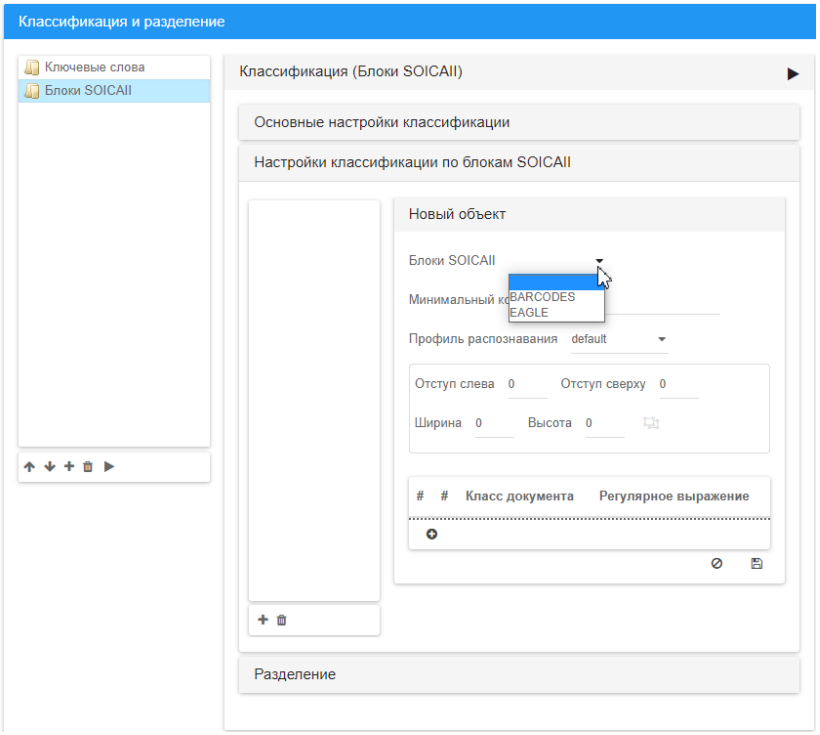
Результирующий класс – В (в том случае, если 27% больше минимального процента для классификации страницы)

Параметр «Размер свертки» - Это размер изображения, в пикселях которое получается из исходного изображения перед вычислением цветовой классификации. Т е предварительно происходит уменьшение изображение до определенного размера в пикселях и только после этого вычисляется контраст, яркость и насыщенность полученных пикселей.

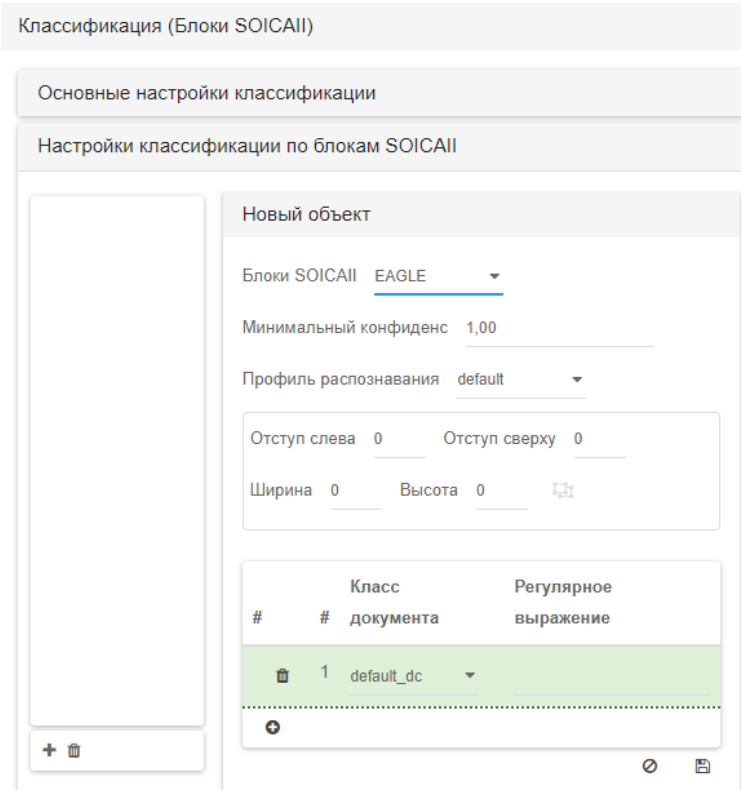


(Рис. 51 Классификация по цвету)

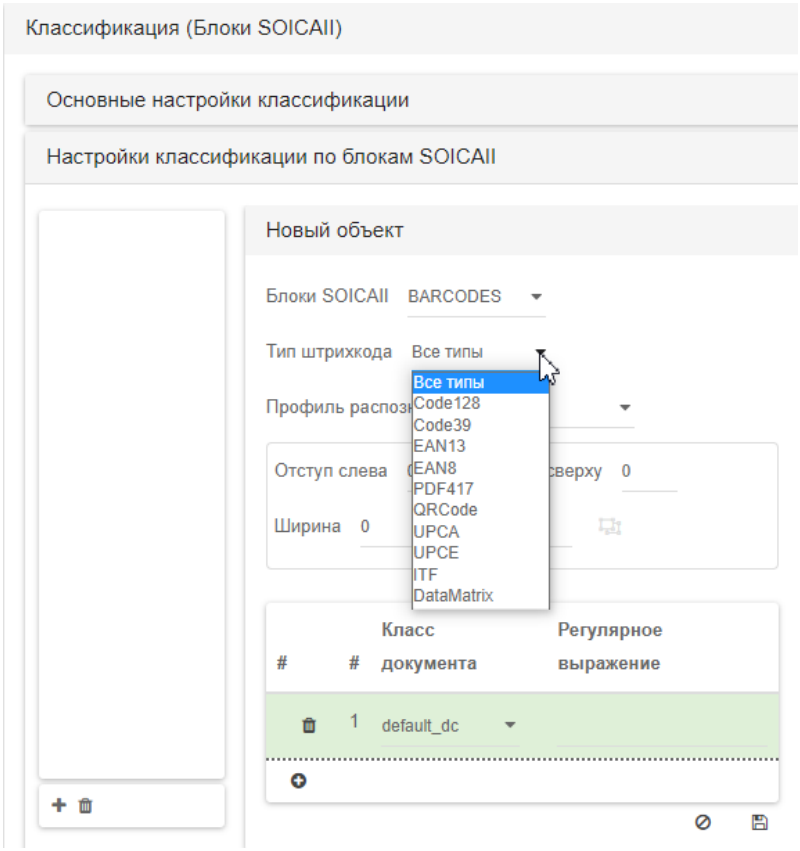
Классификация по Блокам SOICA



В данном случае выполняется поиск штрих кода указанного типа или гербового орла в указанной области указанной репрезентации. Кандидат в результаты классификации рассчитываются присваивается из указанного класса при соответствии найденного штрих кода или гербового орла указанному регулярному выражению.



(Рис. 48.1 Классификация по гербовым орлам)



(Рис. 48.2 Классификация по штрих-коду)

Штрих коды. Список штрих кодов для поиска в указанной области указанной репрезентации.

Тип штрих кода. Указывает формат штрих кода, которые необходимо найти на изображении страницы. Варианты форматов: "Все типы", "Code128", "Code39", "EAN13", "EAN8", "PDF417", "QRCode", "UPCA", "UPCE", "ITF", "DataMatrix".

(Рис. 49 Форматы Штрих-кодов)

Профиль распознавания. Указывает репрезентацию (результаты профиля распознавания) в которых будет осуществлен поиск штрих кодов.

Настройки области. Описывает область репрезентации, в которой будет выполняться поиск штрих кодов. Настройки включают: X, Y, Width, Height.

- **Отступ слева** – отступ от левой границы изображения репрезентации выраженный в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.
- **Отступ сверху** – отступ от верхней границы изображения репрезентации выраженный в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.
- **Ширина** – ширина области, выраженная в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.
- **Высота** – высота области, выраженная в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.

Суммы Отступ слева, Ширина и Отступ сверху, Высота так же не могут быть больше 100.

Влияние на класс. Список классов документа на кандидаты которых будет влиять указанный штрих код.

Класс документа. Указывает класс документа, значение кандидата которого будет изменено на 100 найденным штрих кодом, в случае соответствия им регулярного выражения.

Регулярное выражение. Указывает регулярное выражение, при совпадении найденного штрих кода, с которым, кандидату указанного класса будет присвоена степень доверия 100.

Дополнительные способы классификации.

В системе предусмотрен способ упорядочивания страниц одного документа в нужном порядке.

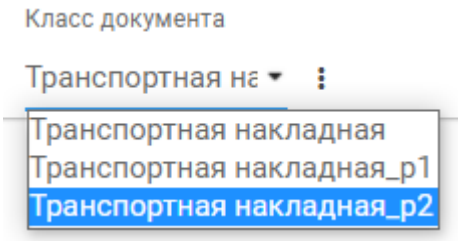
Для этого необходимо создать общий класс документа и классы для каждой страницы документа. Класс_документа_p1, Класс_документа_p2 и так далее.

После настройки нахождения каждой страницы и их классификации к соответствующему типу из них автоматически формируется единый документ общего класса с правильным порядком страниц.

Пример 12:

На импорт поступает файл с ограниченным числом страниц. Это Транспортная накладная из двух страниц, но страницы могут быть перепутаны местами.

В системе создается класс «Транспортная накладна» и классы страниц «Транспортная накладная_p1» и «Транспортная накладная_p2».

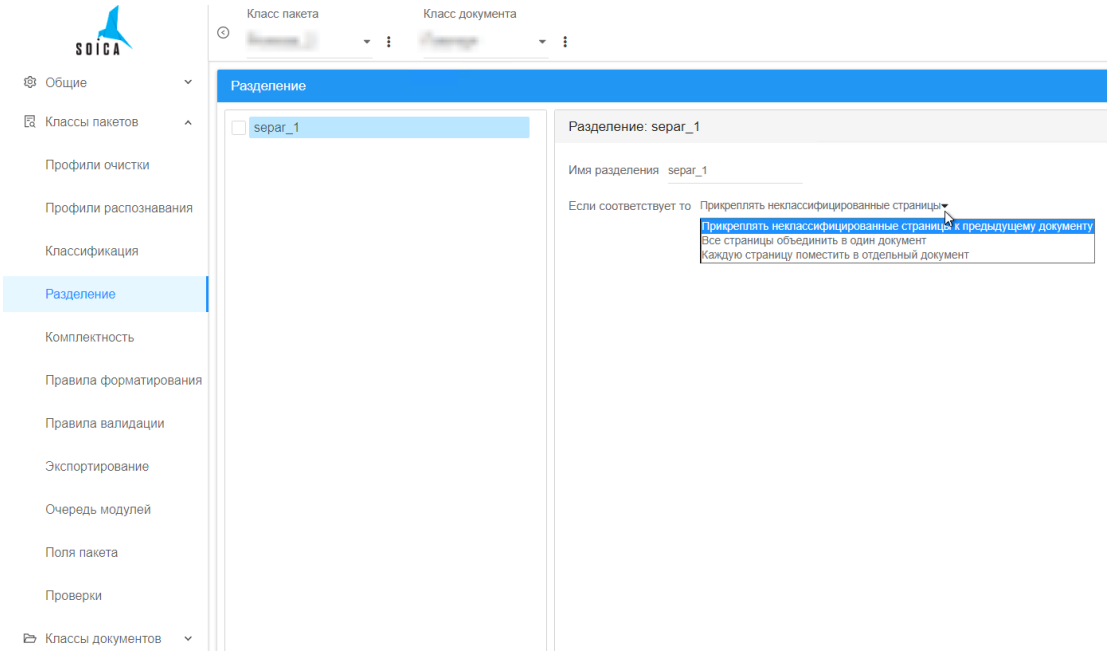


Первая и вторая страница поступившего файла перепутаны местами. Первая страница классифицируется как «Транспортная накладная_p2», а вторая страница как «Транспортная накладная_p1» по ключевым словам.

После этого два классифицированных документа объединяются в один документ «Транспортная накладная» и упорядочиваются в соответствии с классификацией. После этого происходят извлечение данных с документа с правильно отсортированными страницами.

2.3.4 Разделение.

Разделение - это процесс формирования документов из страниц пакета по результатам классификации и указанному алгоритму.



(Рис. 52 Разделение страниц)

Алгоритмы создания документов указывает принцип, по которому будут формироваться документы. Варианты:

— **Все страницы в один документ**. В данном случае будет создан один документ, класс его будет взят из результатов классификации первой страницы пакета.
— **Каждую страницу в отдельный документ**. В данном случае из каждой страницы пакета будет создан отдельный документ с классом этой страницы.
— **Прикреплять неклассифицированные страницы**. В данном случае классифицированные страницы (или первая страница, если даже он не классифицирована) создают документы с классом этих страниц, а

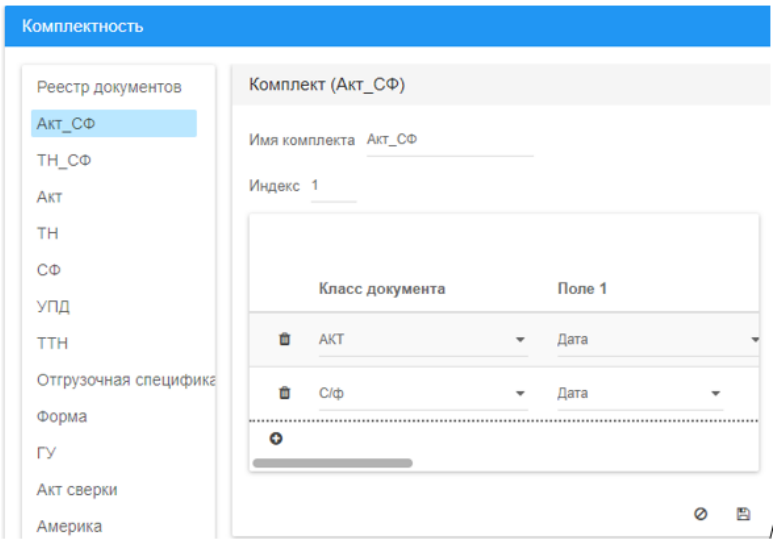
неклассифицированные страницы прикрепляются к документам в которых находятся предыдущие страницы.

Можно создать несколько алгоритмов создания документов по разным принципам формирования.

2.3.5 Комплектность

Функционал комплектности необходим для выделения указанного числа документов из одного пакета (родительский пакет) в отдельный самостоятельный пакет (комплект). Проверка комплектности происходит в рамках этапа распознавания, после получения полей и таблиц.

Комплектность может происходить автоматически в процессе распознавания, либо оператором в модуле валидации. Комплектность собирается только из классифицированных документов в одном пакете.



Комплектность может быть осуществлена по совпадению данных в одном или нескольких полях.

Условия комплектности – это сравнение полученных полей разных документов между собой.

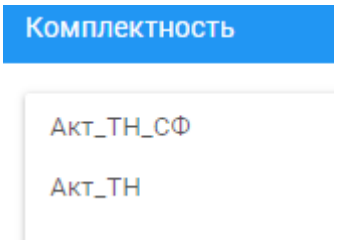
Очередность условий формирует порядок проверки документов на комплект.

То есть на данном примере сначала СФ будет сравниваться с Актом и в случае, если не нашлось подходящего Акта, будет происходить сравнение с ТН.

Комплект может собираться из любого количества документов (Акт_СФ_ТН_Счет, Акт_СФ, ТН_Счет, УПД_Акт_ТН, Паспорт_Снилс и т. д.)

Пример 1:

В комплектности настроены условия в следующей последовательности: Акт_ТН_СФ и Акт_ТН.



На импорт поступили 5 документов: 2 Акта, 2 СФ и 1 ТН. Создался родительский пакет со всеми документами.

После классификации и извлечения данных с каждого документа, проверяется условие Акт_ТН_СФ.

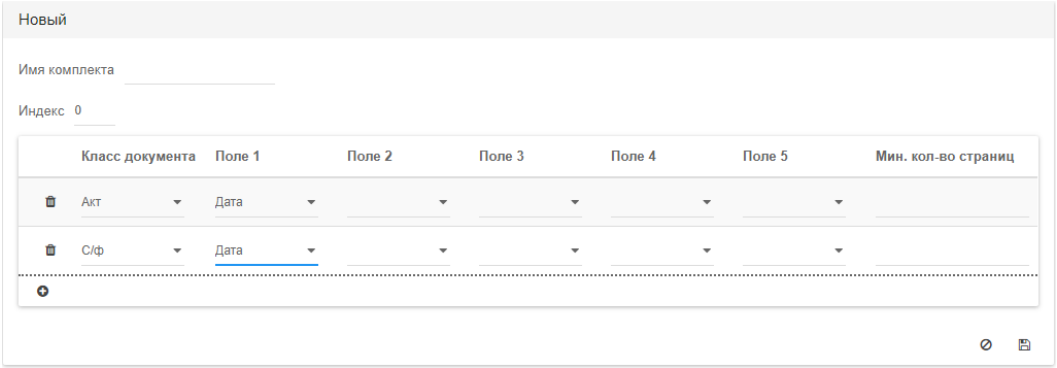
Поля в Акте и ТН совпадают, а в СФ нет. Комплект не собрался по первому условию.

Происходит проверка по условию Акт_ТН.

Поля в одном из актов и ТН совпадают. Эти два документа образуют комплект и образуют отдельный пакет из 2 документов. Остальные документы остаются в родительском пакете.

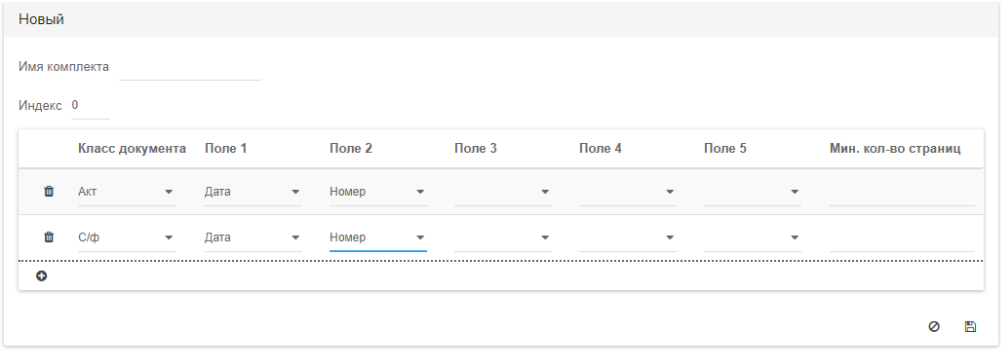
В сборе комплекта могут участвовать любые настроенные документы.

Пример сравнения по одному полю:



В данном случае сравниваются поле «Дата» из документа «Акт» и поле «Дата» из документа «С/ф».

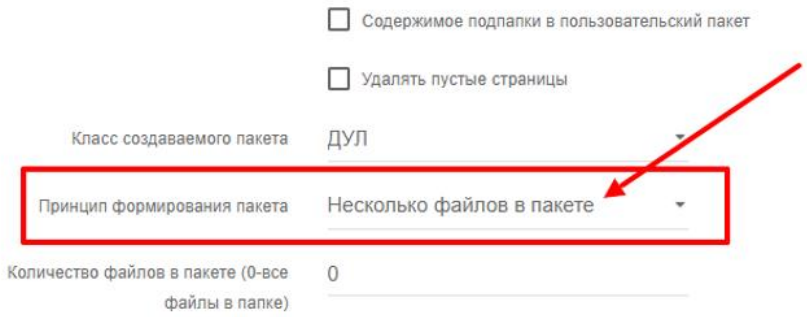
Пример сравнения по двум полям:



В данном случае сравниваются поля «Дата» и «Номер» из документа «Акт» и поле «Дата» и «Номер» из документа «С/ф».

Если сравниваемые поля удовлетворяют условию, то создается отдельный пакет и отправляется по сценарию обработки родительского класса пакета. В имя нового пакета можно поместить название поля-условия, либо системные данные (дата создания пакета, логин пользователя, запустившего процесс обработки и так далее), либо данные, распознанные на документе. В новый пакет добавляются документы со всеми страницами.

Важно! Проверка осуществляется только внутри одного пакета. Т.е в настройках сценария Импорта в поле «Принцип формирования пакета» должно быть выбрано условие «Несколько файлов в пакете»



Лишние документы в родительском пакете.

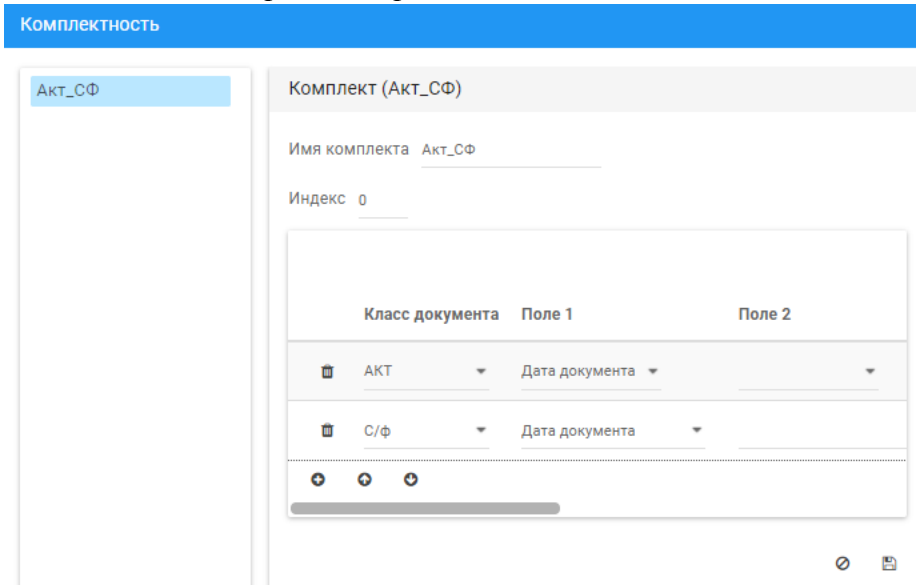
Если в родительском пакете остаются документы, не удовлетворяющие условиям комплектности, то можно настроить два варианта:

1. Документы, не удовлетворяющие условиям, переходят в «Модуль контроля качества». В этом модуле все документы, не собравшиеся в комплект, объединяются в отдельный пакет и ожидают проверки оператора в модуле валидации, либо поступления документа, который подойдет по настроенным условиям.
2. Документы, не удовлетворяющие условиям, остаются в родительском пакете и ожидают проверки оператора в модуле валидации.

Важно! При появлении в родительском пакете двух одинаковых документов, подходящих под настроенные условия, соберется один комплект с одним из одинаковых документов. Документ-дубль будет помещен в «Модуль контроля качества», либо останется в родительском пакете.

Пример 2:

Условия комплектности: С/ф и Акт при совпадении даты.

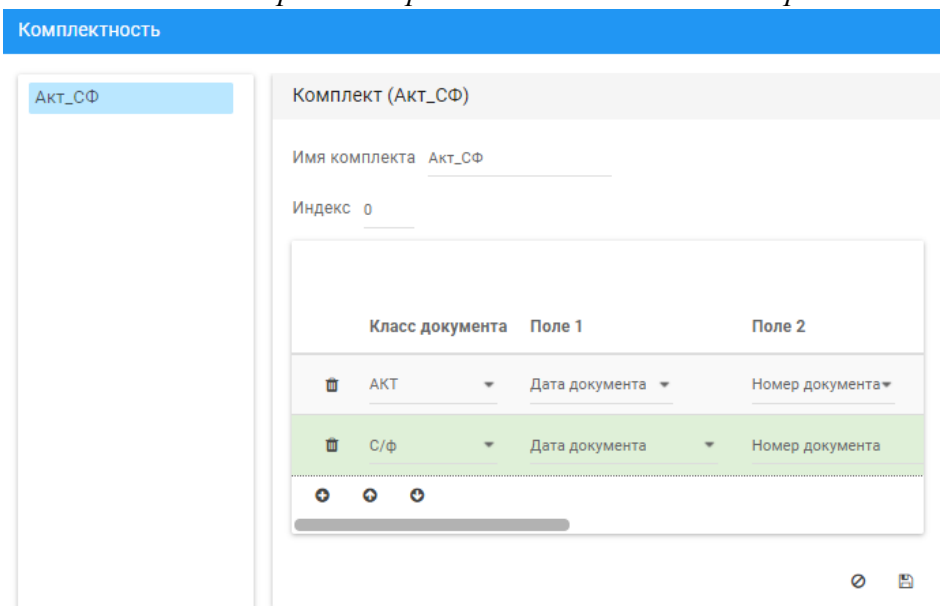


На импорте в родительском пакете два акта с подходящей датой к С/ф.

При сборе комплектности будет выделен самостоятельный пакет с одной С/ф и одним Актом, который поступил на импорт первым. Документ-дубль останется в родительском пакете, либо перейдет в «Модуль контроля качества».

Пример 3:

Условия комплектности: С/ф и Акт при совпадении даты и номера.



На импорте в родительском пакете два акта с подходящим номером, но только один из них с подходящей датой.

При сборе комплектности будет выделен самостоятельный пакет с одной С/ф и одним Актом, в которых присутствует полное совпадение и выполнены условия комплектности, то есть оба поля совпадают. Второй акт останется в родительском пакете для проверки оператором валидации, либо перейдет в «Модуль контроля качества».

Функционал выделения нового пакета документа относительно документа.

Система может формировать комплект относительно найденного документа в пакете.

То есть при обработке потока документов при нахождении указанного класса все остальные документы выделяется в новый пакет.

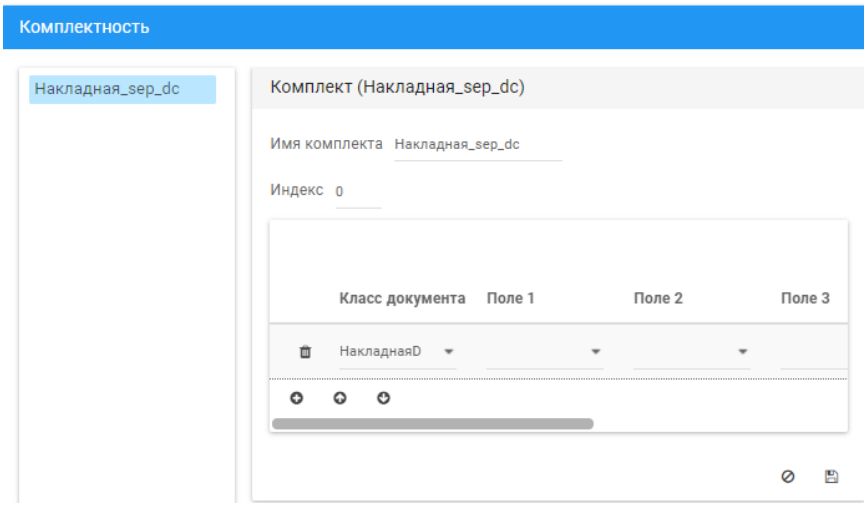
Для этого необходимо в имени комплекта добавить «sep_dc» и указать класс пакета для разделения.

Пример 4:

На импорте поток документов из набора Накладная, Счет, Акт. Этот набор повторяется в указанной последовательности.

Необходимо чтобы при нахождении накладной последующие за ней акт и счет собирались в комплект: Накладная_Счет_Акт. Где накладная индикатор нового пакета.

Настраиваем комплектность с ключом sep_dc:



Теперь при нахождении в родительском пакете документа «НакладнаяD» все последующие за ней документы, до новой накладной, будут выделены в самостоятельный пакет.

Создание дополнительного документа-дубля для сбора комплекта.

В системе предусмотрено создание документа дубля из имеющегося классифицированного документа. Для этого необходимо классифицировать документы и передать в извлекаемое поле всю область изображения.

Затем в меню переклассификации выбрать это поле и назначить ему отдельный класс.

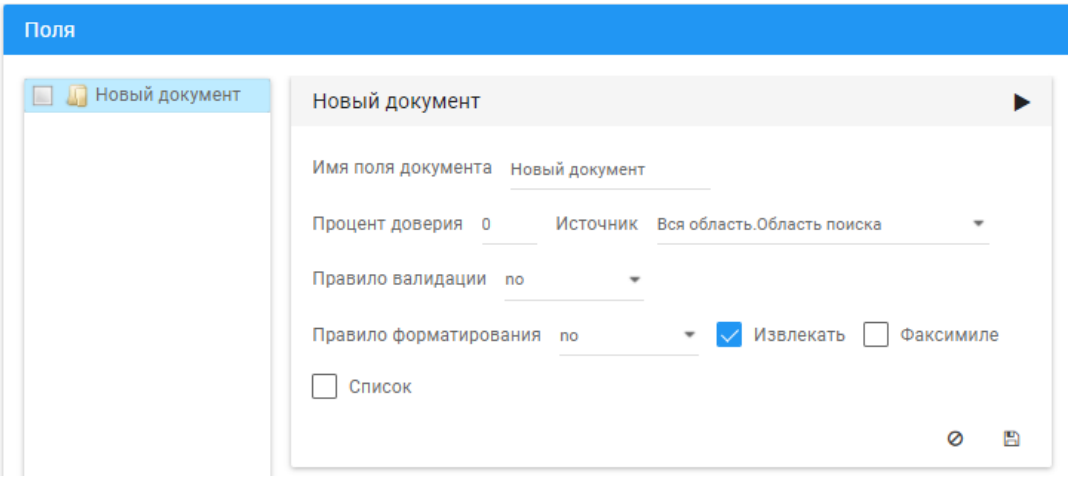
Система автоматически скопирует все изображение и присвоит ему указанный класс. Затем извлечет настроенные данные в соответствии с присвоенным классом.

Пример 5:

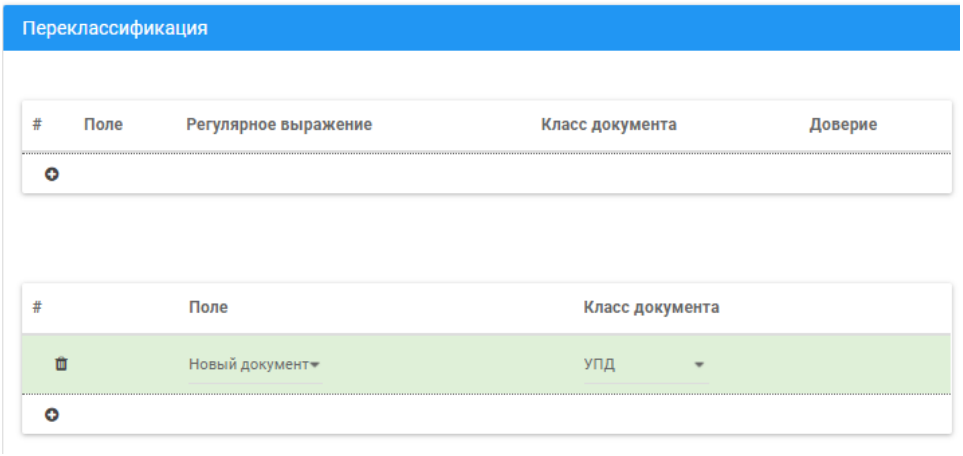
На импорт поступили документы: 1 УПД и 2 Акта.

Необходимо собрать два комплекта: УПД и Акт 1; УПД и Акт 2.

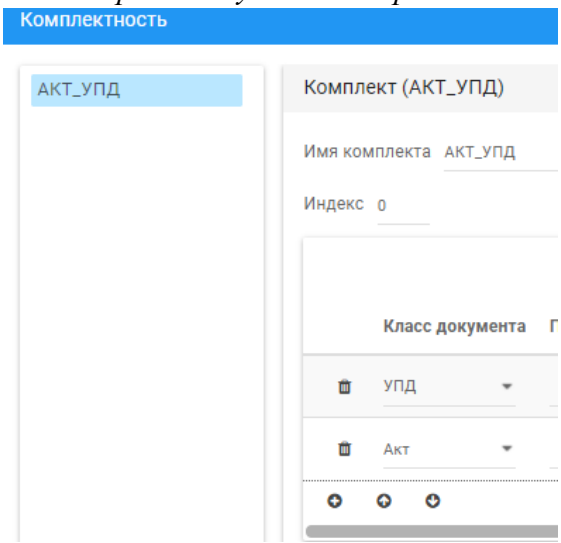
В классе УПД настраиваем поле «Новый документ» и передаем туда область всего изображения УПД.



В меню переклассификация выбираем созданное поле и присваеваемый класс:



В меню комплектности настраиваем условие сбора комплекта Акт_УПД



Теперь в процессе обработке система первично классифицирует доккументы как: УПД, Акт, Акт.

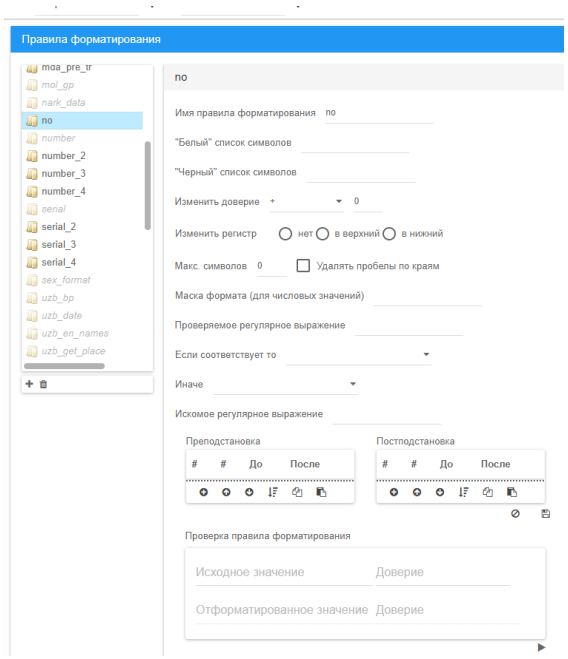
Затем создаст из указанной области-поля документ-дубль с классом УПД.

После проверки условий комплектности выделится два самостоятельных пакета: УПД_Акт и УПД (документ-дубль)_Акт.

2.3.6 Правила форматирования.

Форматирование – процесс изменения текста в соответствии с заданными правилами. Правила форматирования используется для любых полученных значений.

Область правил форматирования делится на два раздела: Список правил форматирования и Настройка выбранного правила.



(Рис. 53 Настройка правила форматирования)

Белый список символов. Указывает список тех символов, которые останутся после этого этапа форматирования.

Черный список символов. Указывает список символов, которые будут удалены на этом этапе форматирования.

Изменять регистр. Указывает будет ли изменен регистр букв из текста на верхний или нижний.

Удалять пробелы по краям. Если эта опция выбрана, то пробелы в начале или конце форматируемого текста будут удалены.

Максимальное количество символов. Если оно больше 0, то на этом этапе длина текста уменьшается до указанного значения, если она была больше.

Проверяемое регулярное выражение. Указывает регулярное выражение, которому должен соответствовать текст после форматирования, чтобы выполнялось правило форматирования указанное для случая соответствия регулярному выражению. Иначе выполняется правило форматирования указанное для случая несоответствия регулярному выражению.

Если соответствует, то. Указывает правило форматирования указанное для случая соответствия регулярному выражению.

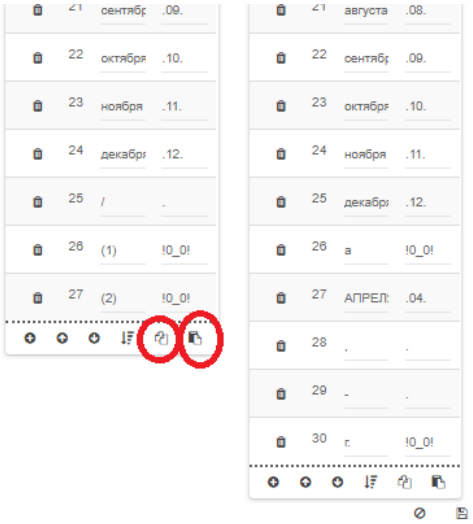
Иначе. Указывает правило форматирования указанное для случая несоответствия регулярному выражению, которое также выполняется, если регулярное выражение не указано.

Тип изменения доверия. Указывает будет ли увеличиваться, уменьшаться или приравняется степень доверия указанному значению.

Величина изменения доверия. Указывает величину изменения доверия. Диапазон: -100 – 100.

Чтобы не изменять значение степени доверия можно выбрать «увеличение на 0». Изменение доверия может применяться для альтернатив, полей и ячеек таблицы.

Список предпостановки и постподстановки. Указывает пары, состоящие из совокупностей символов: «до» и «после». Если в форматируемом тексте есть совокупность символов из столбца «до», то она заменяется на совокупность символов из столбца после. Замена происходит в указанной последовательности. Предпостановка выполняется первым этапом форматирования, постподстановка – последним. Так же вместо текста может быть указана конструкция вида: !N_M!, где N и M – числа. Есть возможность копировать и вставлять таблицу пред- и постподстановки из уже имеющегося правила форматирования.



(Рис. 53.1 Кнопки копирования и вставки правил форматирования)

Если эта конструкция применяется в столбце «До», то это можно читать как: «Заметить M символов исходной строки с позиции N». Если N больше длины строки, то текст из «После» будет добавлен в конец строки. Если M=0, то будет произведена вставка текста из «После» после символа номер N. Если эта конструкция применяется в столбце «После», то часть исходной строки будет использоваться в качестве заменителя. Если M=0, тогда символы из «До», будут удаляться. Пример как получить дату из формата: «ГГММДД» в «ДД.ММ.ГГ»:

Исходный текст: 851112

До	После	Промежуточный результат
!6_0!	!0_2!	85111285
!0_2!	!0_0!	111285
!0_0!	!2_2!	12111285
!4_2!	!0_0!	121185
!2_0!	.	12.1185
!5_0!	.	12.11.85

Пример:

Так же существует конструкция для замены/удаления части строки относительно ключевого текста:

!>текст!> - находит все после указанного текста;

!<текст!< - находит все до указанного текста.

В первом случае берется часть строки после указанного текста, причем, если этот текст будет несколько раз в строке, то будет использован последний случай. Т.е., например:

Строка: к*ор***о**навирус

Правило подстановки: !>о!>

Результат: навирус

В случае со второй конструкцией также берется последнее вхождение, но уже справа, и часть строки берется перед этим текстом:

Строка: самоизол**я**ци*я*

Правило подстановки: !<я!<

Результат: самоизол

Новый вид правил: !>>*текст*!>> и !<<*текст*!<<

В этом случае берется, наоборот – первое вхождение. Т.е.:

Строка: к**о***ро*навирус

Правило подстановки: !>о!>

Результат: ронавирус

Строка: самоизоляци**я**

Правило подстановки: !<я!<

Результат: самоизоляци

Конструкция **#conf#** - берет конфиденс значения, т.е. если написано ДО - !0_0!, ПОСЛЕ - **#conf#**, то строка «пандемия» с конфиденсом 50, станет – «50пандемия»

Конструкция **~regex~** - выполняет поиск в строке первое вхождение указанного регулярного выражения, т.е.:

~\d{4}~ в строке «первый случай заболевания произошел в 2019 году», результатом будет «2019»

!numstring! – позволяет представить число в виде строки, т.е. если строка была «1866654», то результатом будет: «один миллион восемьсот шестьдесят шесть тысяч шестьсот пятьдесят четыре»

!memM! – если стоит в ДО, то запоминает в переменную с номером M (от 0 до 9), то что стоит ПОСЛЕ, если же стоит ПОСЛЕ, то считывает из переменной

Пример перевода суммы в строковое представление:

#	#	До	После
🗑	1	!mem0!	!>,!>
🗑	2	!>,!>	!0_0!
🗑	3	,	!0_0!
🗑	4	!0_1000!	!numstring!
🗑	5		
🗑	6	!1000_0!	рублей
🗑	7	один рублей	один рубль
🗑	8	два рублей	два рубля
🗑	9	три рублей	три рубля
🗑	10	четыре рублей	четыре рубля
🗑	11	рублей	рублей
🗑	12	!1000_0!	!mem0!
🗑	13	!1000_0!	копеек
🗑	14	1 копеек	1 копейка
🗑	15	2 копеек	2 копейки
🗑	16	3 копеек	3 копейки
🗑	17	4 копеек	4 копейки

Строка: «123456,78»

Результат: «Сто двадцать три тысячи четыреста пятьдесят шесть рублей 78 копеек»

Для удаления необходимо в столбце «после» указать конструкцию: !0_0!. Если необходимо заменить или дописать какой-то текст после или до указанного выражения в столбце «После» необходимо указать новое выражение. При удалении сам текст остается в строке.

Исходный текст: 851112

До	После	Промежуточный результат
!>11!>	!0_0!	8511
!<11!<	!0_0!	1112
!>11!>	проверка	8511проверка
!<11!<	проверка	проверка1112
!>11!>	!2_0!	85115
!<11!<	!2_2!	511112

Существуют правила:

@имя_поля@ - берет значение поля документа. Можно использовать только в полях или таблицах, а не в локаторах, т.к. локаторы выполняются до полей. Причем поле, имя которого указано должно быть получено до того поля или таблицы, где это правило применено.

!calcM! , где M – количество знаков после запятой. Вычисляет формулу, если текст является формулой. Т.е. исходный текст: «(10+3)*2» в результате дает «26».

!datefromms1970! – преобразует количество милисекунд от 00:00 01.01.1970 в читаемую дату.

В пред(пост)подстановке в форматторе можно использовать поиск по внешнему источнику.

Конструкция**: *\D+_\d+_\d+_((bw)|(all))_\d+(=.+=)?***

Например: ***источник_1_2_all_70=\d+=***

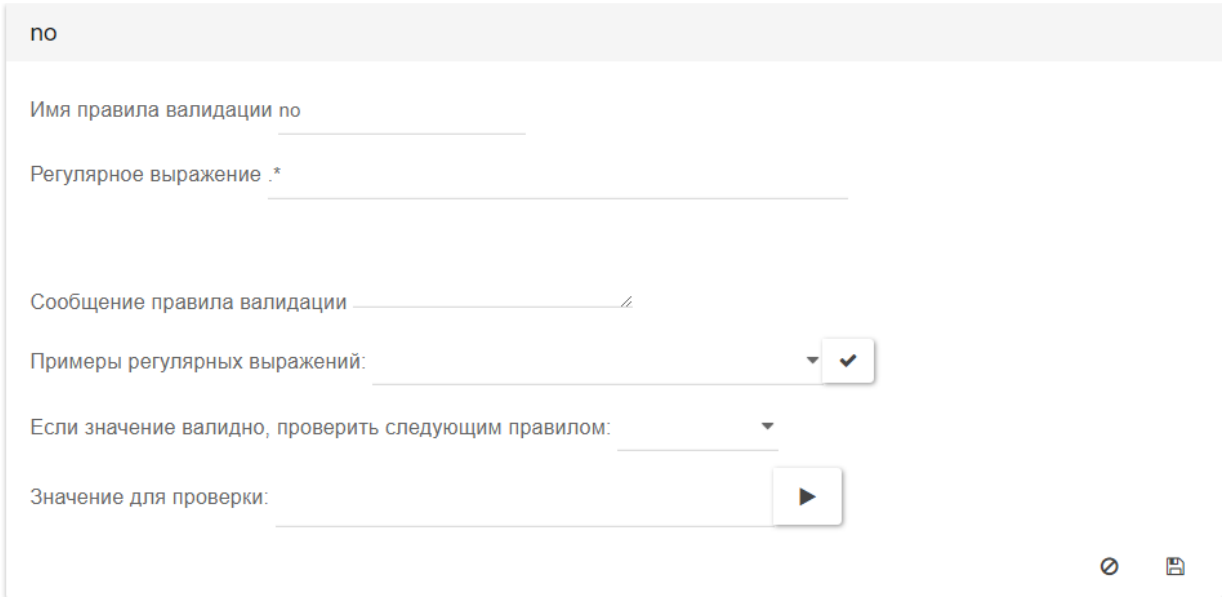
Это значит: используем внешний источник «источник», сравниваем по неточному соответствию всю входную строку с текстом 1го столбца таблицы, предварительного обработанного регулярным выражением “\d+”, и если результат сравнения больше 70%, возвращаем текст из 2го столбца строки

Если вместо “**all**“ напишем «**bw**», то сравнение будет происходить не всех входящей строки, а, по ее словам, по очереди, до тех пор, пока результат сравнения не будет удовлетворять условиям.

Регулярное выражения для преобразования текста из БД перед сравнением можно опустить, тогда сравнение будет происходить с неизмененным тестом.

2.3.7 Правила валидации.

Валидация процесс проверки допустимости содержимого поля или ячейки таблицы. Правила валидации не позволяют оператору подтвердить некорректные данные для отправки в целевую систему. Область правил валидации делится на два раздела: Список правил валидации и Настройка выбранного правила.



(Рис. 54 Настройка правила валидации)

Регулярное выражение. Указывает регулярное выражение на совпадение, с которым проверяется текст поля или ячейки таблицы.

Сообщение правила валидации. Указывает текст, который записывается в качестве сообщения для поля или ячейки таблицы, если совпадения с регулярным выражением не произошло.

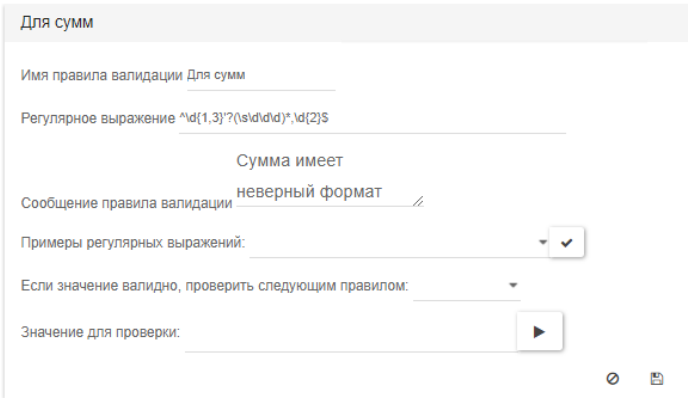
Примеры регулярных выражений. Примеры хранятся в базе на сервере. В выпадающем списке представлены самые популярные регулярные выражения.

Следующее правило валидации. Указывает правило валидации, которое проверяется, если текущее правило выдало положительный результат.

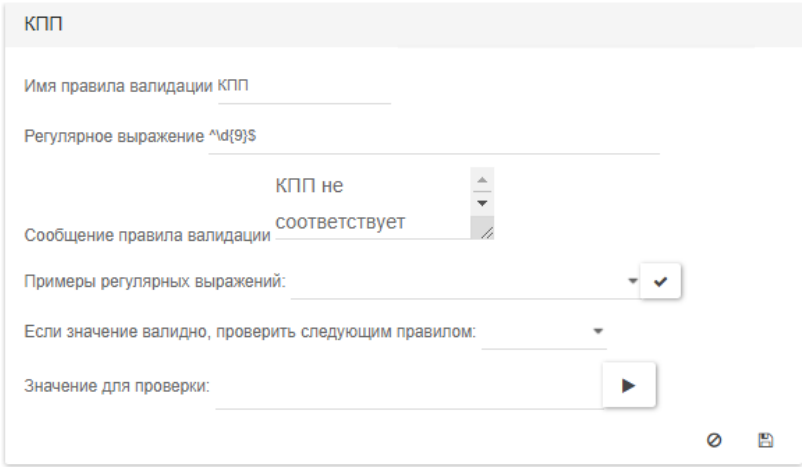
Значение для проверки. В это поле можно написать текстовую строку и проверить работу настроенного правила валидации нажав кнопку 

Примеры:

Правило Валидации для проверки сумм



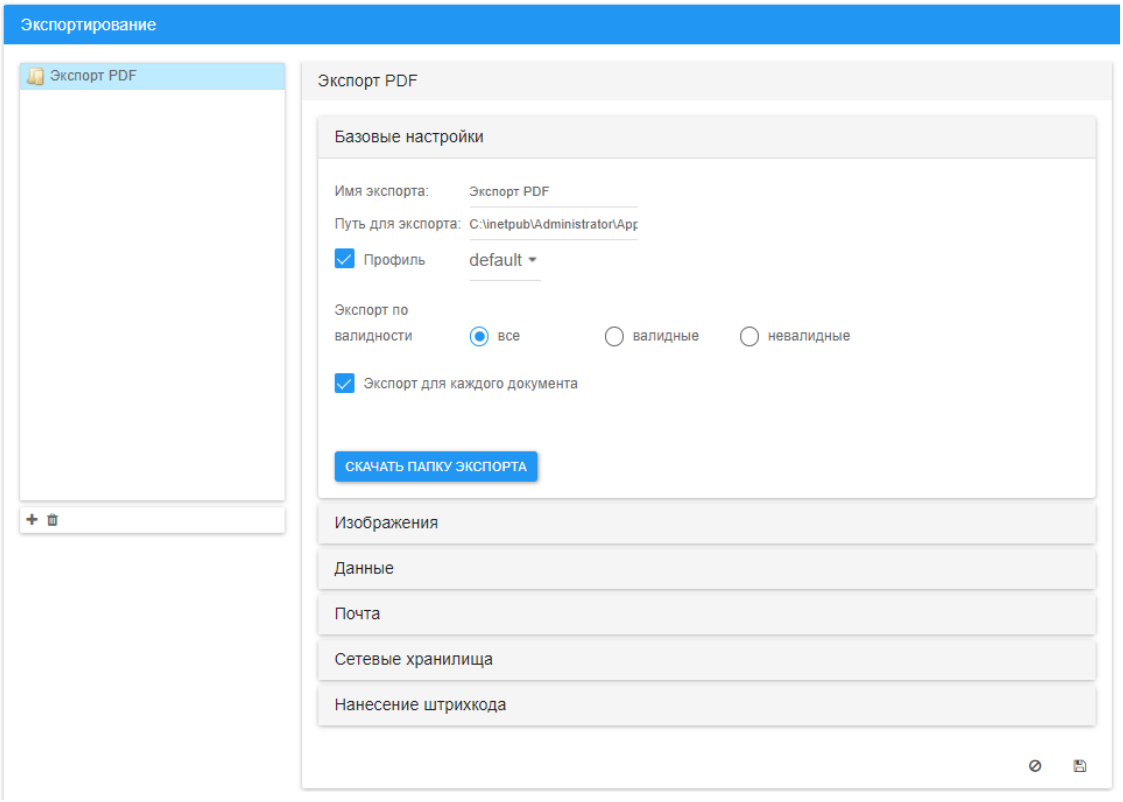
Правило валидации для проверки правильности КПП



2.3.8 Экспортирование.

Экспортирование – процесс вывода найденных данных в заданном виде. (Подробнее в пункте 4. Экспорт)

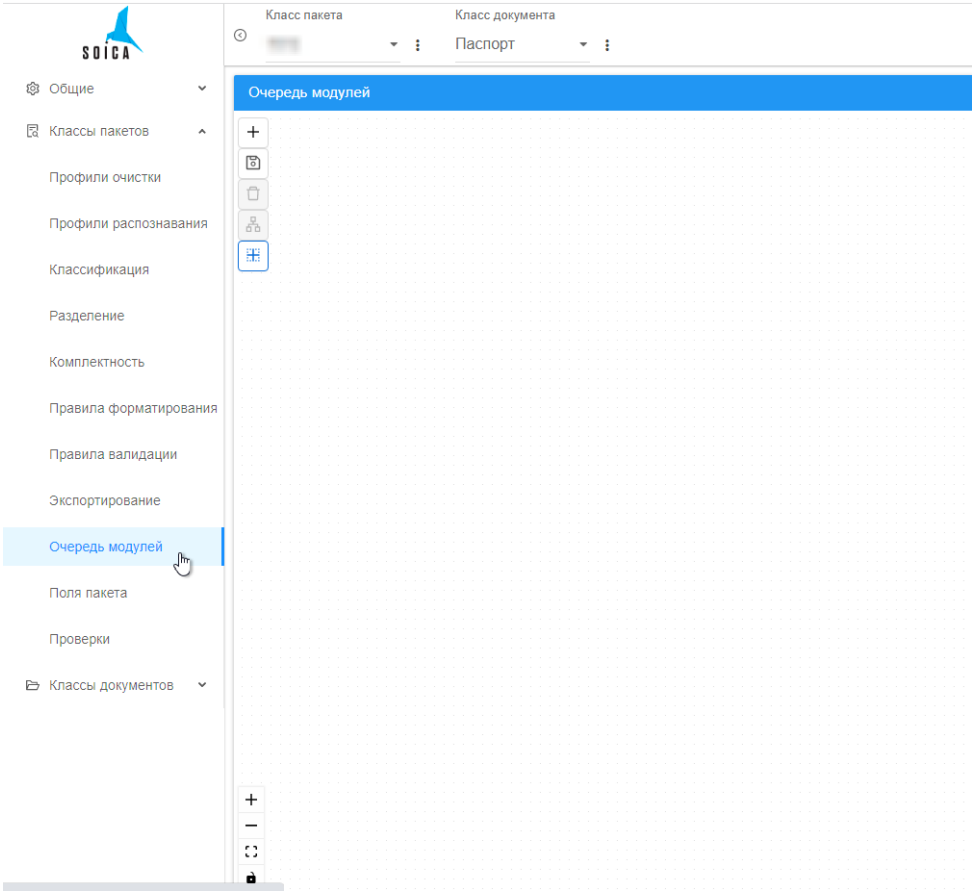
Общий вид интерфейса:



(Рис. 55 Интерфейс настройки экспорта)

2.3.9 Очередь модулей.

Очередь модулей предназначена для создания сценария передвижения пакета по модулям в виде блок схемы со связями (воркфлоу схемы). Каждый тип модуля можно использовать несколько раз, а также наложить условия для следования пакета по определенной ветке обработки в зависимости от выполнения условий на определенном этапе обработки пакета.



(Интерфейс раздела Очередь модулей)

 - Добавить элемент.

 - Сохранить изменения.

 - Удалить выделенный элемент.

 - Сбросить связь к начальному виду.

 - Привязка к сетке.

- Выбор размера сетки, по которой можно перемещать узлы. Размер сетки влияет на отображение линий связи.

 - Приблизить изображение.

 - Уменьшить изображение.

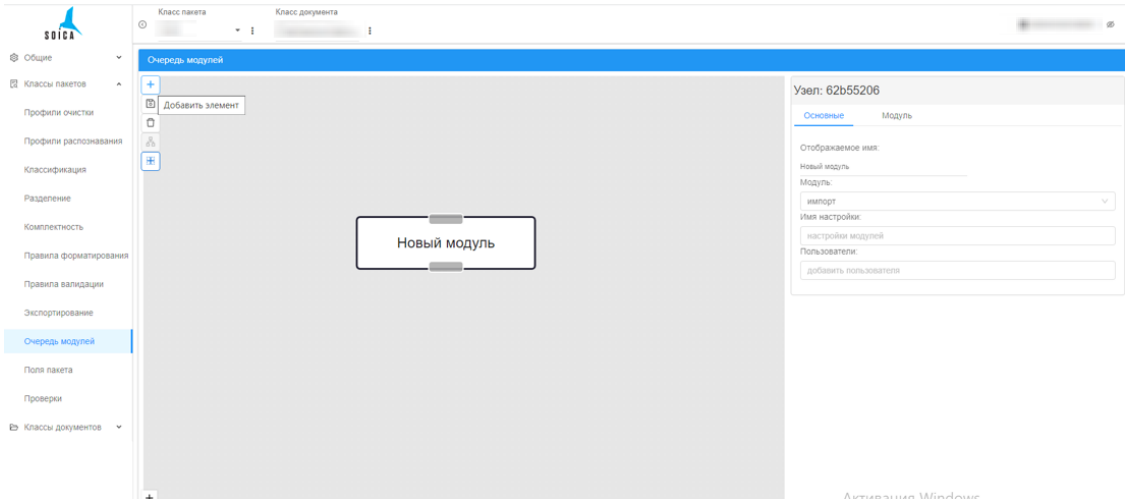
 - Подходящий вид. Позволяет отобразить всю схему.

 - Переключить интерактивность.

2.3.10 Создание и настройка модулей схемы.

С помощью кнопки «Добавить элемент» необходимо добавить нужное количество модулей, которое будет перемещаться пакет в процессе обработки. Далее необходимо

указать связи между модулями, наложить проверки на связи, в случаях их необходимости (такие связи станут синим цветом) и сохранить изменения.

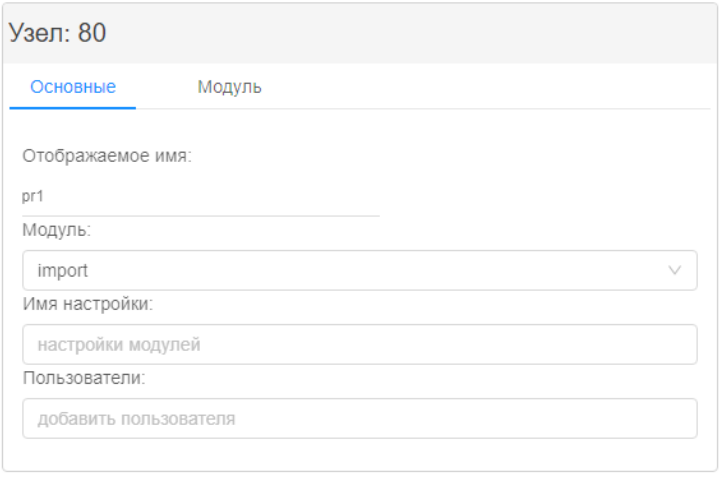


(Добавление нового элемента)

Необходимо выполнить настройку для каждого модуля в правой панели настроек. Обязательно необходимо указать имя и тип модуля. Можно выбрать следующие модули: import, export, classification, documentExtract, separate, complect, validation, afterextractvalidate.

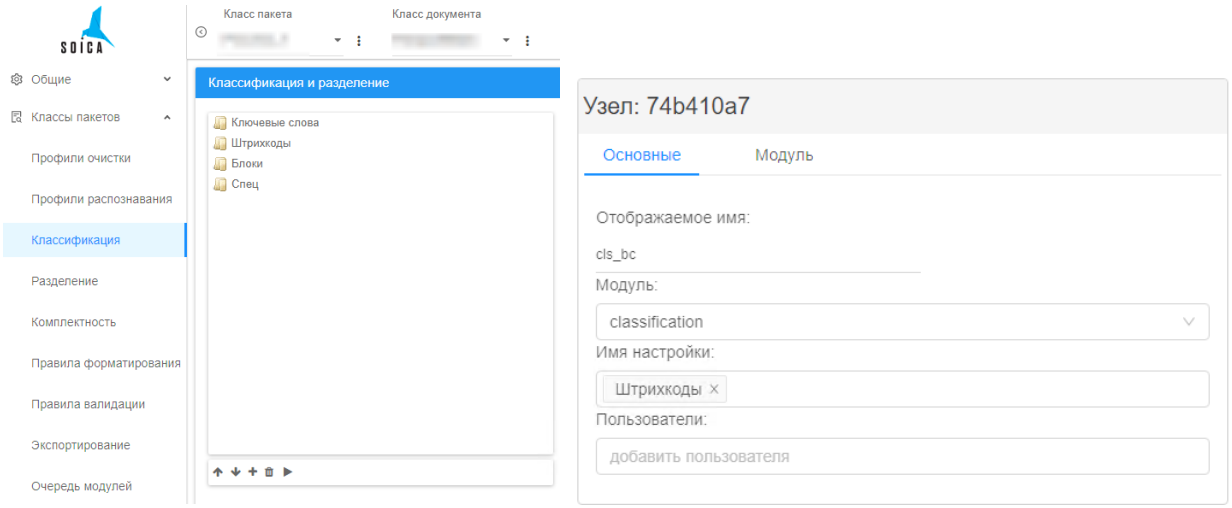
В зависимости от выбранного типа модуля необходимо выполнить дополнительные настройки в разделах Основные и/или Модуль.

ВАЖНО! Для модуля import не нужно производить дополнительные настройки в правой панели. Сценарий импорта создаётся для данного Класса пакета (проекта) в разделе Способы обработки (Общие) и автоматически привязывается к модулю при выборе типа import в настройках.



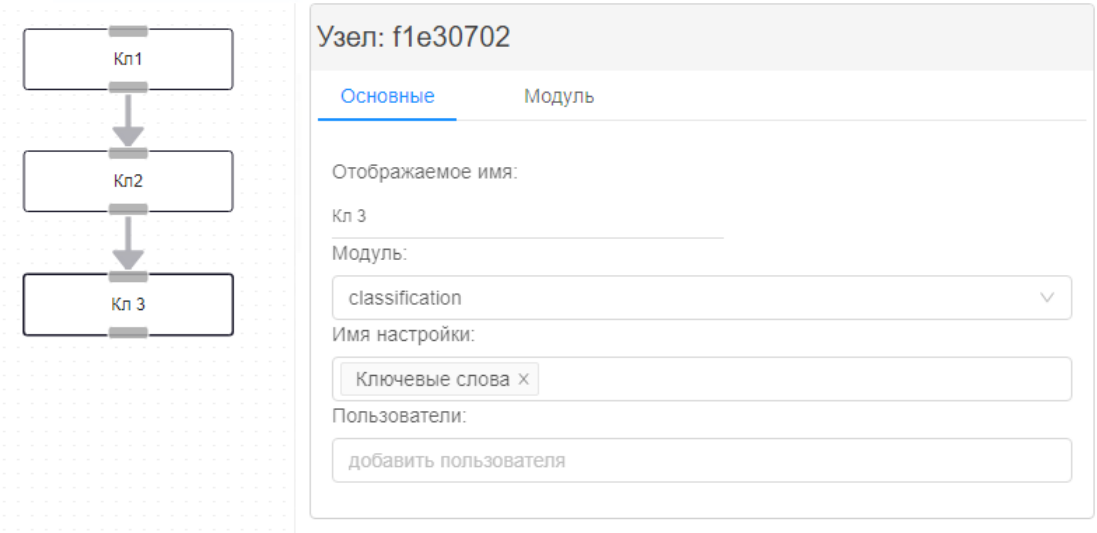
(Панель настроек модуля import)

В поле Имя настройки для модуля classification необходимо выбрать один вид классификации, которая уже настроена в разделе Классификация.



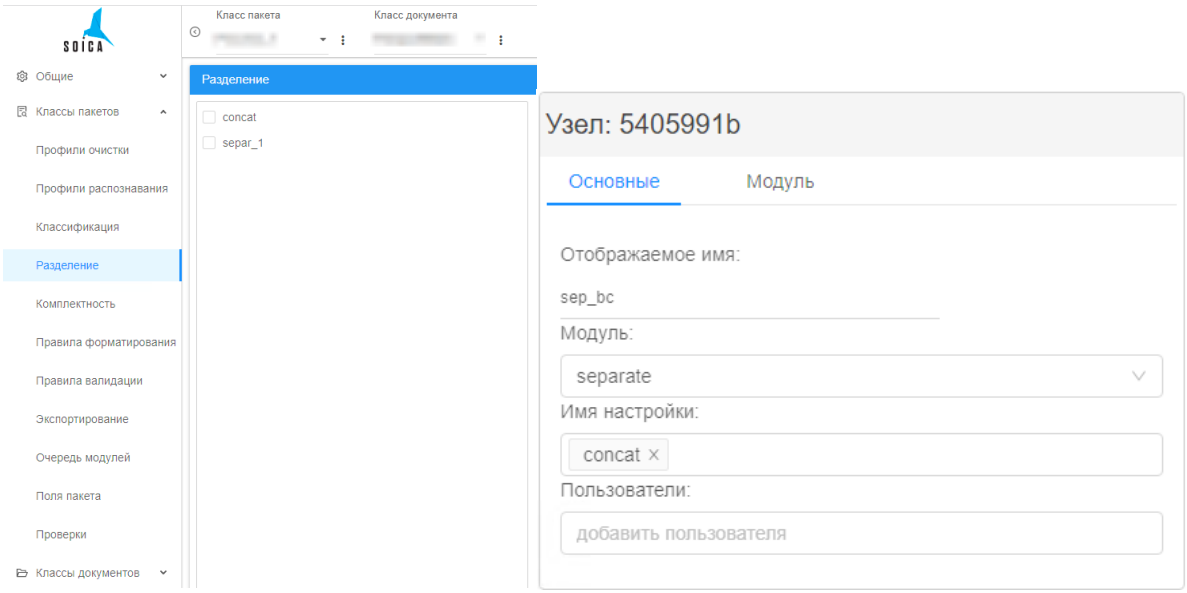
(Панель настроек модуля classification)

Если в процессе обработки пакета необходимо выполнить несколько видов/этапов классификации, необходимо создать новый модуль для каждого вида классификации. В поле Имя настройки в каждом модуле выбрать только один тип классификации. Далее необходимо расположить модули в нужном порядке и связать между собой.



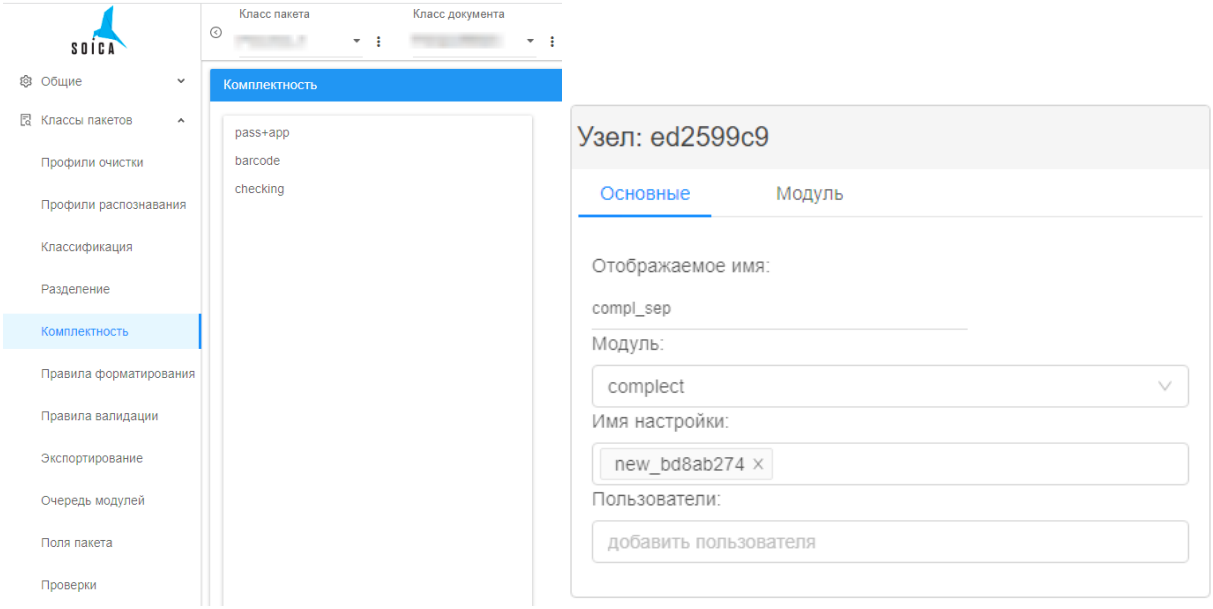
(Пример настройка нескольких этапов классификации)

В поле Имя настройки для модуля separate необходимо выбрать один из ранее созданных в разделе Разделение методов разделения. Если в процессе обработки пакета необходимо выполнить несколько этапов разделения, необходимо создать для каждого этапа новый модуль, в настройках каждого выбрать только один метод разделения и создать связи между модулями.



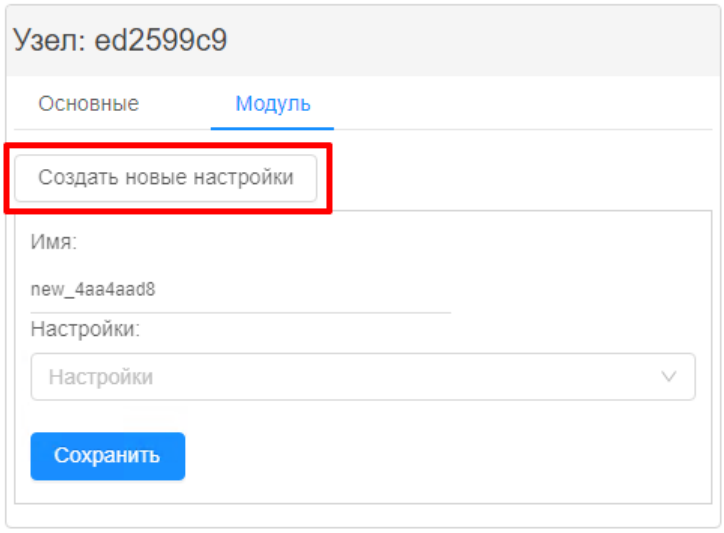
(Панель настроек модуля separate)

В поле Имя настройки модуля complect необходимо выбрать ранее созданный вид комплектности.



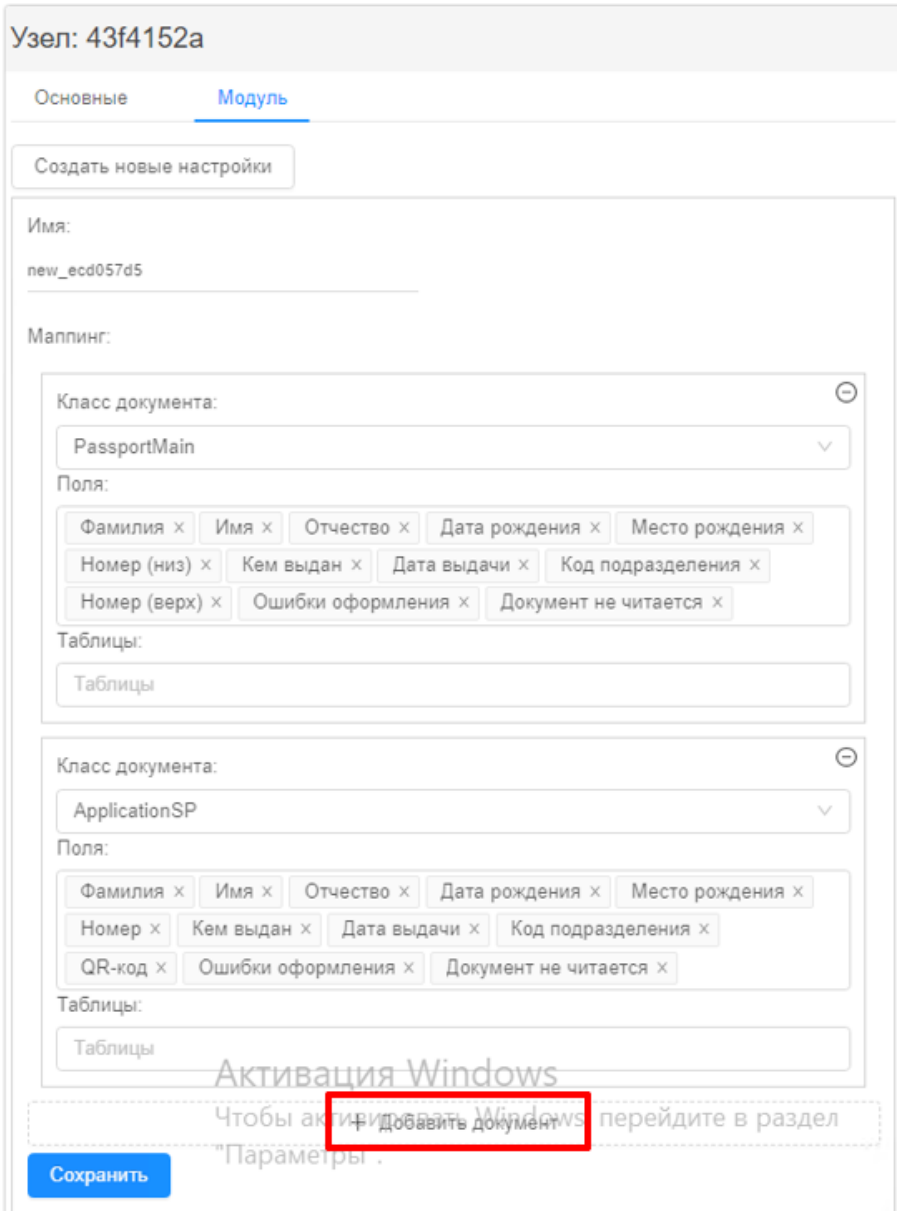
(Панель настроек модуля complect)

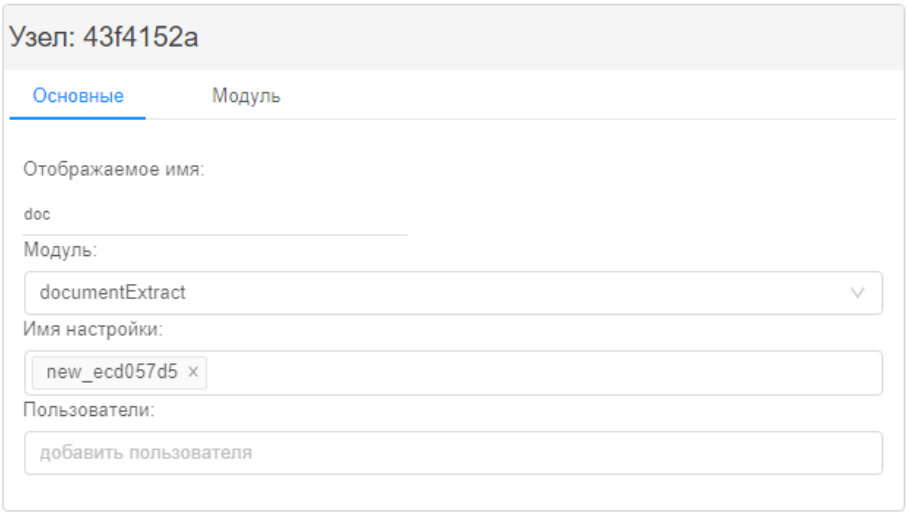
В настройках данного модуля есть возможность создать новые настройки комплектности во вкладке Модуль с помощью соответствующей кнопки.



(Создание новых настроек комплектности)

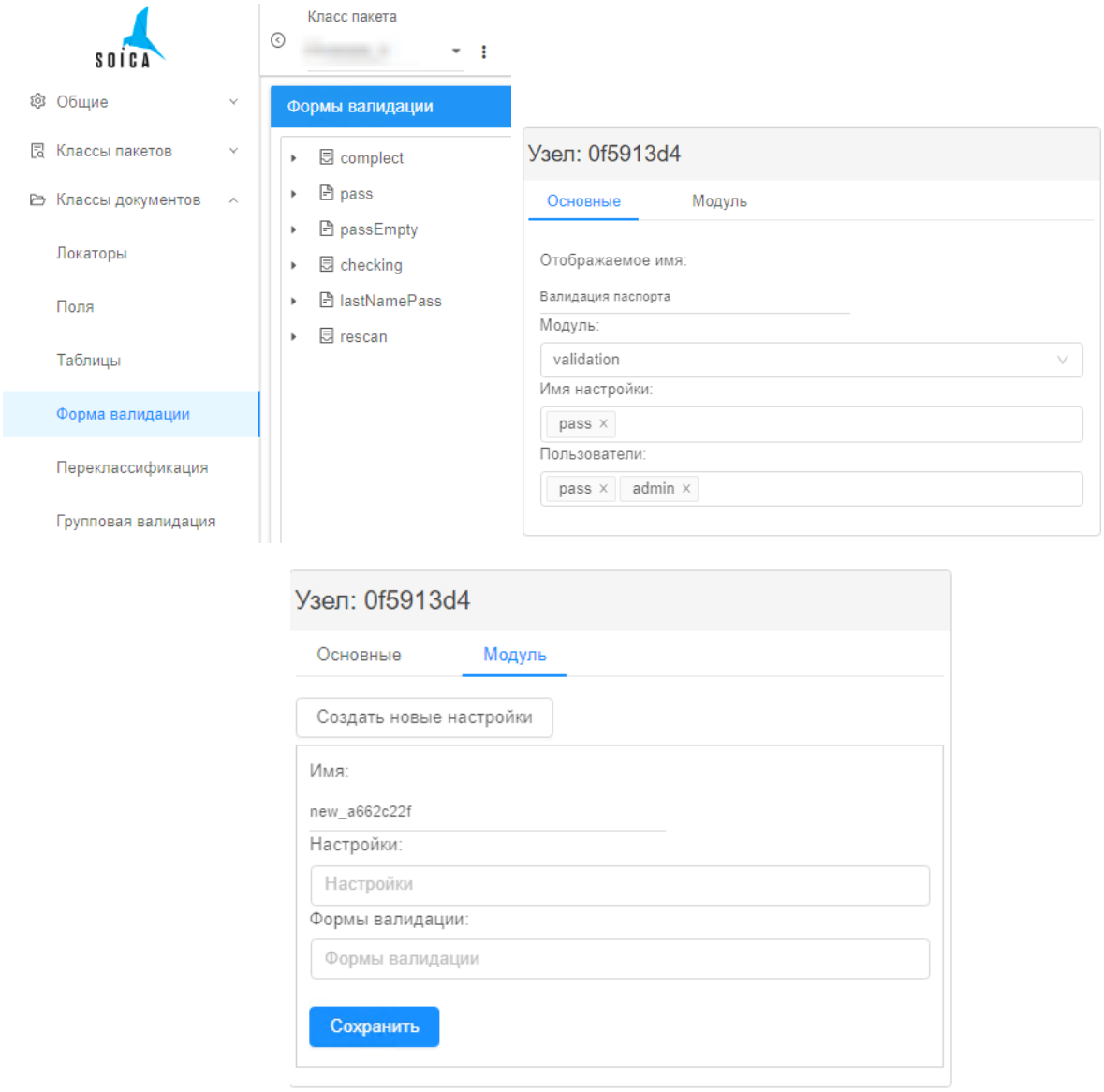
В настройках модуля documentExtract необходимо выполнить настройку в двух вкладках. Сначала необходимо создать новые настройки в разделе Модуль. Для этого необходимо указать имя и настроить Маппинг, т.е. выбрать из уже созданных класс документов, поля и таблицы для извлечения данных для данного типа документа. Для того, чтобы добавить ещё одни класс документов для маппинга, необходимо нажать кнопку «Добавить документ». После настройки маппинга для всех классов документов (типов документов, на которых извлекаются данные) необходимо нажать кнопку «Сохранить». Далее нужно перейти во вкладку Основные и в поле Имя настройки выбрать созданную во вкладке Модуль настройку с маппингом.





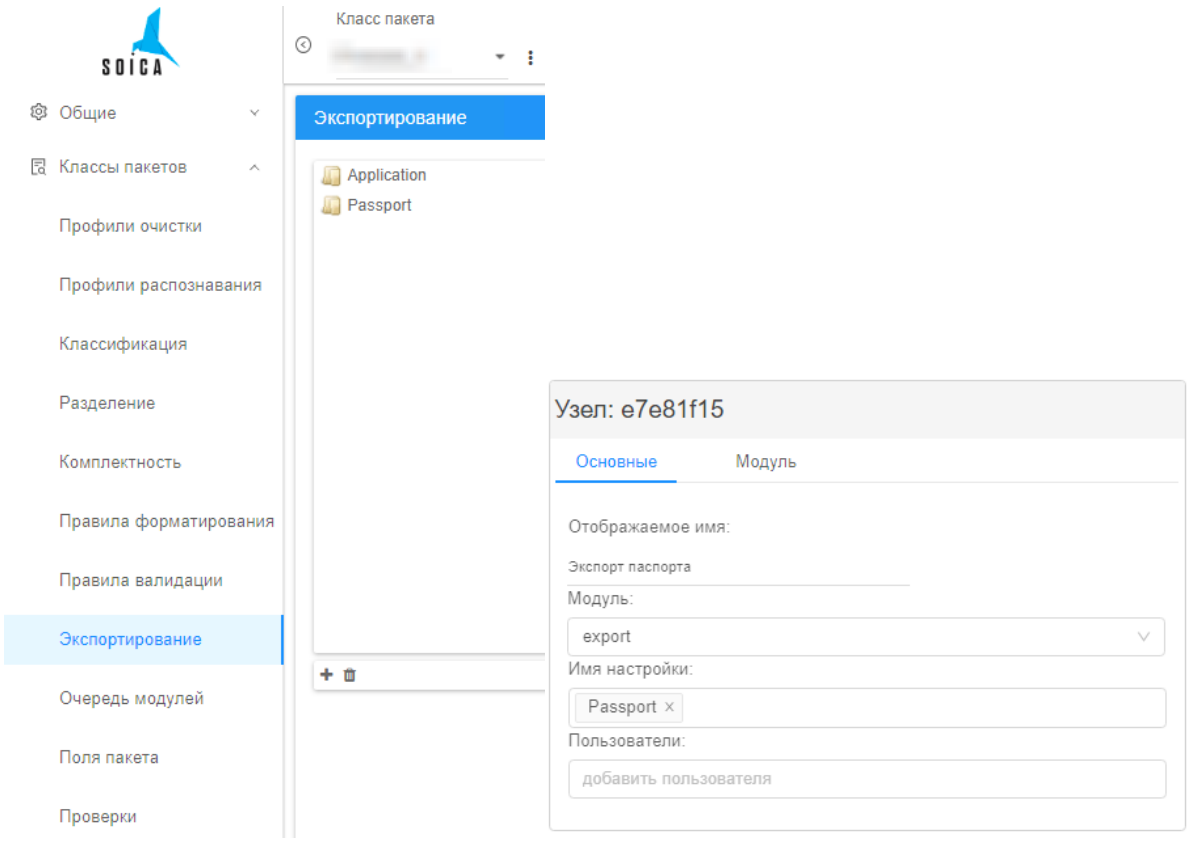
(Настройка модуля documentExtract)

В Основных настройках модуля validation необходимо выбрать имя уже настроенной Формы валидации и пользователей. Есть возможность создать новые настройки во вкладке Модуль.



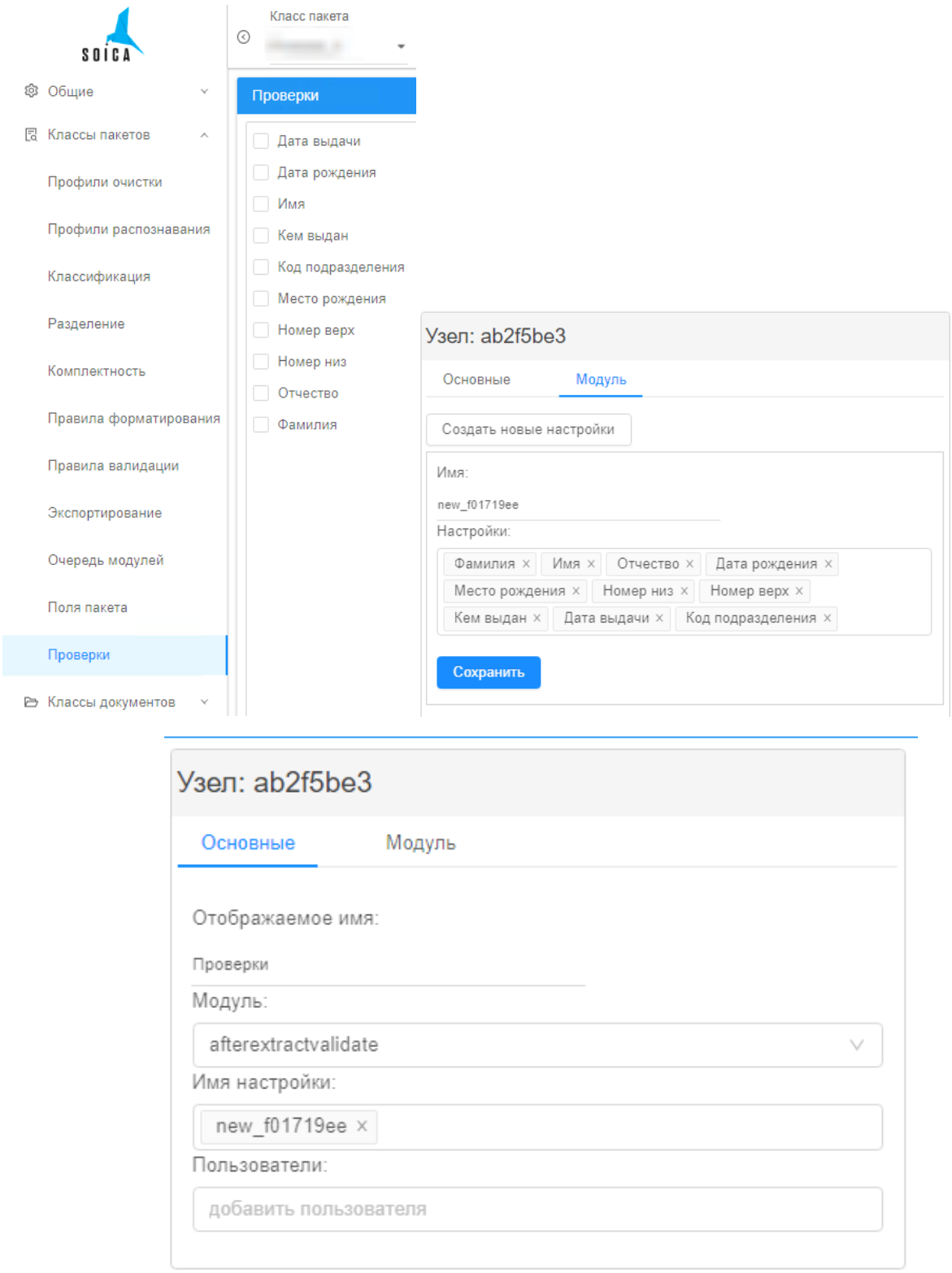
(Настройка модуля validation)

В поле Имя настройки в Основных настройках модуля export необходимо выбрать сценарий экспорта из уже созданных в разделе Экспортирование. Экспортирование – процесс вывода найденных данных в заданном виде.



(Настройка модуля export)

Модуль afterextractvalidate предназначен для выполнения проверок и условий. Настройку данного модуля необходимо выполнить в двух вкладках.

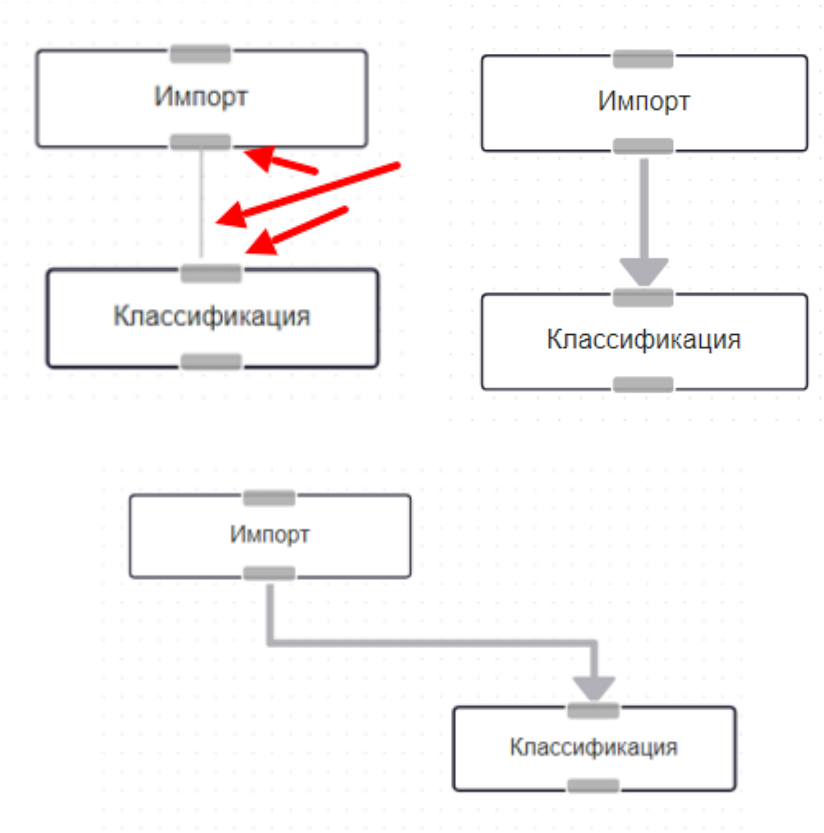


(Настройка модуля afterextractvalidate)

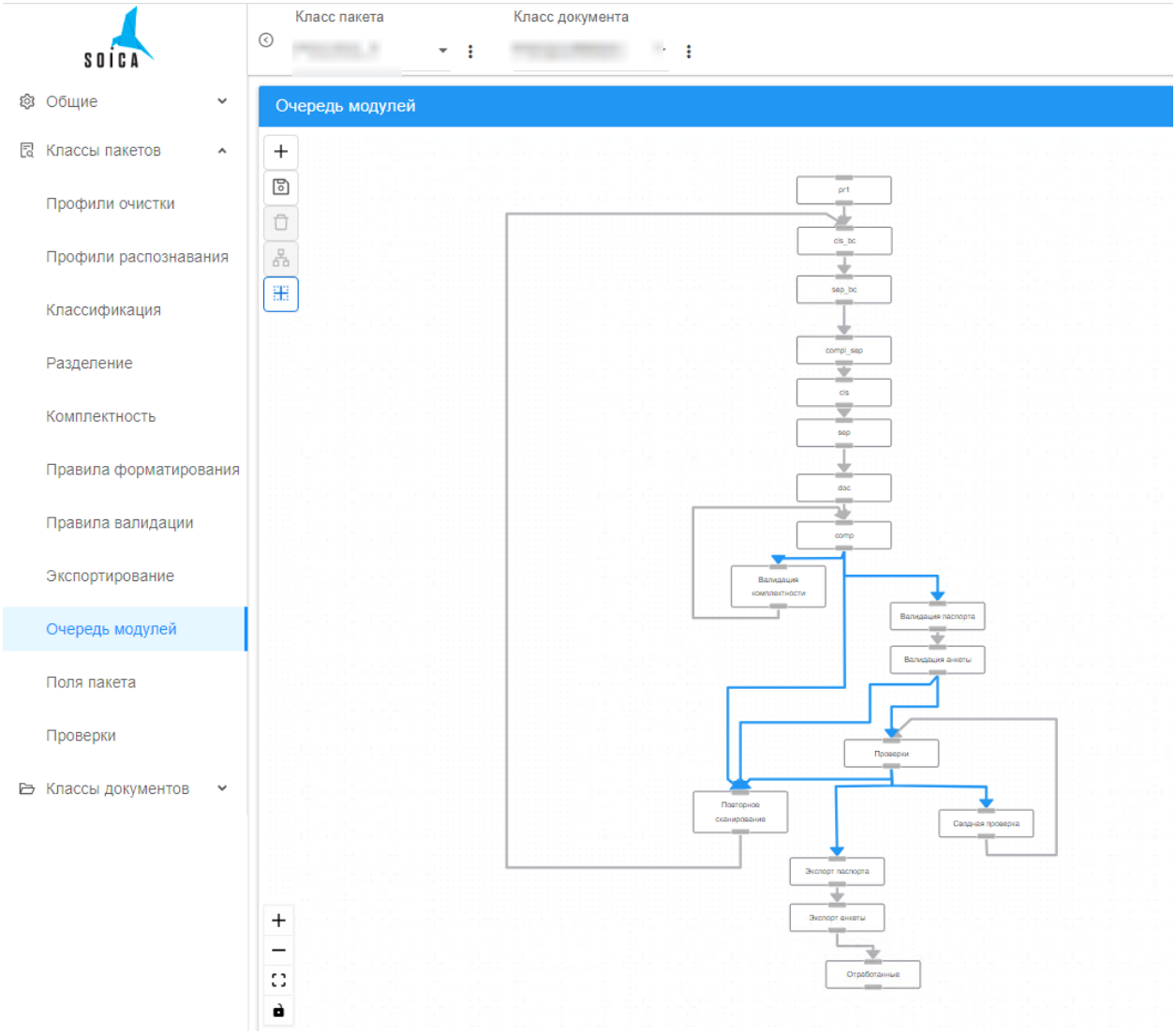
Сначала во вкладке Модуль задается имя настройки, а в поле Настройки выбираются ранее созданные в разделе Проверки условия проверок. Далее во вкладке Основное необходимо задать имя модуля, выбрать тип модуля afterextractvalidate, в поле Имя настройки выбрать имя созданной во вкладке Модуль настройки.

2.3.11 Связи между модулями.

Для того, чтобы выстроить логическую цепочку, по которой будет обрабатываться пакет, необходимо правильно создать связи между модулями. Для того, чтобы создать связь между модулями необходимо встать мышкой на середину модуля, выделенную серым цветом, зажать левую кнопку мыши и провести линию до середины нужного модуля, выделенной серым цветом. Связь между модулями выглядит как стрелка серого или синего цвета. Модули можно свободно перемещать по рабочей области, при этом настроенные связи между модулями будут сами видоизменяться.

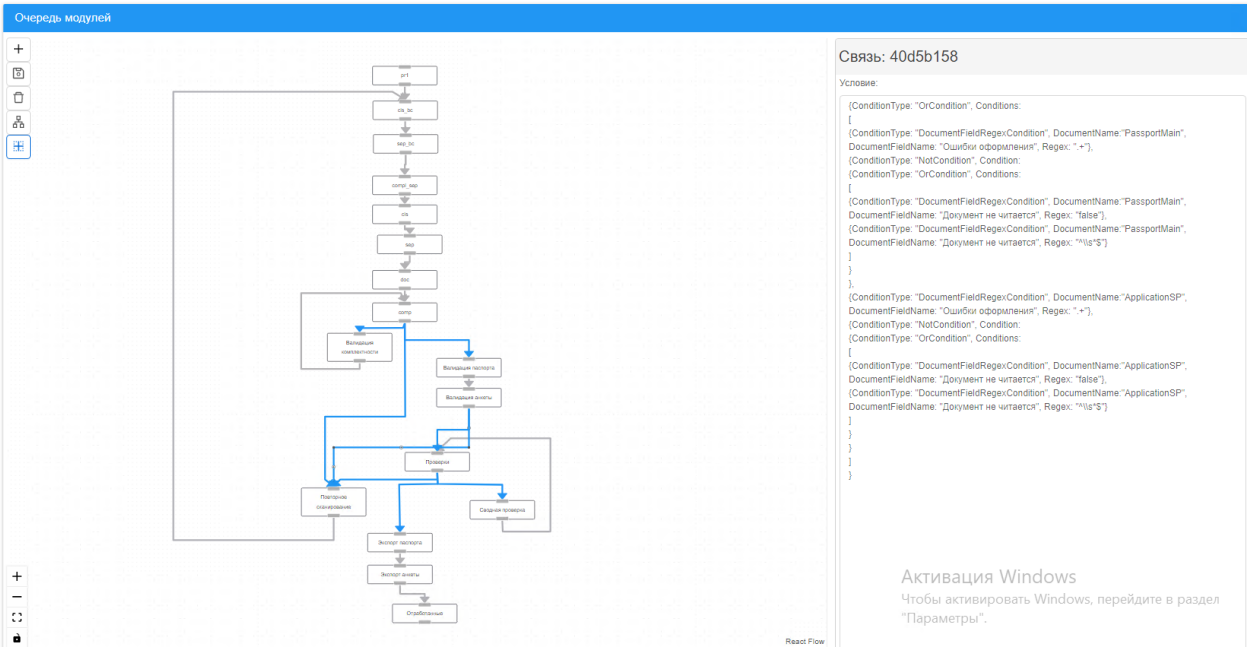


(Настройка связей между модулями)



(Настроенная схема)

Связи между модулями, выделенные синим цветом, обозначают проверку выполнения определенных условий, после которых пакет переходит по одной из связей на следующий модуль, в зависимости от результата выполнения проверки. Условия прохождения пакета по синей связи проверки условий прописывается в синтаксисе формата Json. Условие позволяет проверить регулярное выражение поля пакета или поля документа. Также можно задать условие и/или/не.



(Пример синтаксиса проверки условий, синяя связь между модулями)

Пример синтаксиса условия.

Распишем пример синтаксиса условия: данное условие проверяет пакет на наличие ошибок заполнения в соответствующих полях, если ошибки есть, то условие будет выполнено.

Строчка (1) обозначает логическую операцию «или», т.е. позволяет выбрать одно из нескольких условий. Где «ConditionType» обозначает тип проверки. «OrCondition» - операция «или». В «Conditions» перечисляются все условия.

Первое условие указано в блоке (2). Оно будет считаться выполненным, если поле «Ошибки оформления» класса документа «PassportMain» содержит текст. «DocumentFieldRegexCondition» – проверка регулярного выражения в поле документа. «DocumentName» – имя класса документа. «DocumentFieldName» - имя поля документа. «Regex» - регулярное выражение.

Второе условие указано в блоки (3). Всё это условие обёрнуто инверсией – логической операцией «не». «NotCondition» - логическая операция «не». В «Condition» указывается условие для инвертирования. В качестве этого условия выступает описанный выше «OrCondition». В свою очередь внутри которого указаны два условия по проверки поля документа «Документ не читается» класса документа «PassportMain». Если текст указанного поля содержит «false» или состоит только лишь из пробельных символов, то условие «OrCondition» выдаст результат «истина». И, соответственно, результат «NotCondition» выдаст результат «ложь». Если поле «Документ не читается» содержит текст отличный от выше описанного, то «NotCondition» выдаст результат «истина» и всё условие примет значение «истина».

Третье условие указано в блоке (4). Это условие аналогично условию из блока (2), за исключением того, что выполняется проверка поле в классе документа «ApplicationSP».

Четвертое условие указано в блоке (5). Это условие аналогично условию из блока (3), за исключением того, что выполняется проверка поле в классе документа «ApplicationSP».

```
{
    ConditionType: "OrCondition", Conditions: (1)
    [
        {
            ConditionType: "DocumentFieldRegexCondition",
            DocumentName:"PassportMain", DocumentFieldName:  "Ошибки
            оформления", Regex: ".+"                                    (2)
        },
        {
            ConditionType: "NotCondition", Condition:
            {
                ConditionType: "OrCondition", Conditions:
                [
                {
                    ConditionType: "DocumentFieldRegexCondition",
                    DocumentName:"PassportMain",
                    DocumentFieldName: "Документ не читается",
                    Regex: "false"
                },
                {
                    ConditionType: "DocumentFieldRegexCondition",
                    DocumentName:"PassportMain",
                    DocumentFieldName: "Документ не читается",
                    Regex: "^\\s*$"
                }
                ]
            }                                                           (3)
        },
        {
            ConditionType: "DocumentFieldRegexCondition",
            DocumentName:"ApplicationSP", DocumentFieldName: "Ошибки
            оформления", Regex: ".+"                                    (4)
        },
        {
            ConditionType: "NotCondition", Condition:
                {
                    ConditionType: "OrCondition", Conditions:
```

```
                    [
                    {
                        ConditionType: "DocumentFieldRegexCondition",
                        DocumentName:"ApplicationSP",
                        DocumentFieldName: "Документ не читается",
                        Regex: "false"
                    },
                    {
                        ConditionType: "DocumentFieldRegexCondition",
                        DocumentName:"ApplicationSP",
                        DocumentFieldName: "Документ не читается",
                        Regex: "^\\s*$"
                    }
                    ]
                }
            }
        ]
}
```

(5)

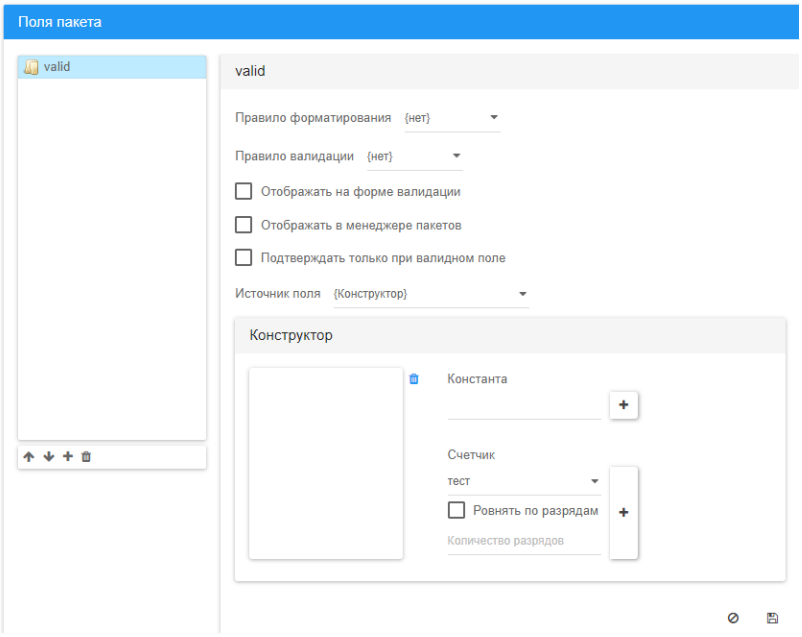
2.3.12 Поля пакета.

Общее описание.

Поля пакета могут содержать информацию, состоящую из констант, заданных в настройках, с использованием инкрементальных счетчиков. Например, каждый новый пакет может содержать поле пакета с номером ящика, из которого достают документы для сканирования: взяли следующий ящик – поменяли поле пакета и обрабатываю дальше. Или же наоборот – номер обработанного пакета за смену: каждый новый пакет +1 к номеру, смена закончилась – счетчик обнулили.

Так же может содержать системные значения: имя файла, отправитель (если пакет пришел по почте), имя пользователя или компьютера, на котором создался пакет и т.д.

Поле пакета с именем valid будет позволять пропускать модуль валидации у валидных пакетов. Валидность – соответствие найденных данных с данными на изображении. Необходимо установить галочку «Отображать на форме импорта».



(Рис. 57 Настройка поля пакета)

Кроме того, поля пакета используются для специальной обработки пакета: настройки кросдокументной и синглдокументой сверки полей задаются в них, так же разработан функционал разбиения пакета на комплекты, их настройки сейчас хранятся так же в полях пакета.

Правило форматировании. Правило форматирования применимое к данному полю.

Привило валидации. Правило валидации применимое к данному полю.

Отображать на форме импорта. Выбранное поле пакета будет доступно на форме настройки сценария импорта.

Отображать в менеджере пакетов. Выбранное поле будет доступно в менеджере пакетов. При выборе данной опции в менеджере пакетов появится столбец «Поля пакета».

Источник поля. Конструктор, либо системное значение такое как имя файла, класс пакета и т д.

Конструктор. Область настройки поля пакета.

Константа. Любой текст в конструкторе не меняющийся при создании полей пакетов.

Счетчик. Значение изменяются относительно заданным настройкам в счетчике при создании поля пакета.

Ровнять по разрядам. Форма счетчика по его разрядности. Т е если в разряде будет 0000, то счетчик начнет отсчет от 0001, 0002… и т д. Так же если разряд 00, то при достижении счетчика значения 100 выводится будет 00, т к чтению счетчика справа на лево.

Расшифровка имён полей пакета:

- Поле с именем **«valid»**. При распознавании оно заполняется в true, если все документы в пакете валидны (все поля и ячейки таблиц валидны), иначе false. Если оно – true, то пакет пропускает модуль валидации.
- Поле с именем «**PROCESS_ROUTE**». Хранит в себе информацию об отражении маршрута в истории пакета.

- Поле содержит в имени **«rcmpl_name»**. В тексте данного поля должно содержаться правило именования пакета при разделении на комплекты следующего вида: Имя_поля1[,Имя_поля2, … Имя_поляN]. Например: ИНН,Номер. В том случае, если такое поле есть, отделяемые комплекты будут именоваться по следующему принципу: Значения_поля1[,Значения_поля2, … Значения_поляN]_Имя_класса1[,Значения_поля1 [,Значения_поля2, … Значения_поляN]_Имя_класса2, … Значения_поля1 [,Значения_поля2, … Значения_поляN]_Имя_классаM…]; Дата Время. Например: 3216549870,154_Акт,1478523690,сф-99882_СФ 05.11.19 12:24.

- Поле содержит в имени **«df»**. В тексте данного поля должно содержаться правило его заполнения из поля документа следующего вида: **Имя_класса1.Имя_поля1[,Имя_класса2.Имя_поля1 … Имя_классаM.Имя_поля1]]. Например: Акт.Контрагент,СФ.Исполнитель.** В указанное поле будет записано первое найденное значение поля.
- Поле содержит в имени **«complect_name»**. При разбиении пакета на комплект поля заполняется именем поля пакета, по правилу из которого собрался комплект.

- Поле содержит в имени **«crosscompare»**. В тексте данного поля должно содержаться правило междокументовой проверки полей следующего вида: **Имя_класса1.Имя_поля1=Имя_класса2.Имя_поля1**. **Например: Акт.Сумма=СФ.Итого.** В случае, если одного из документов нет или больше одного, указанные поля отсутствуют или не равны, то указанные поля будут невалидны.

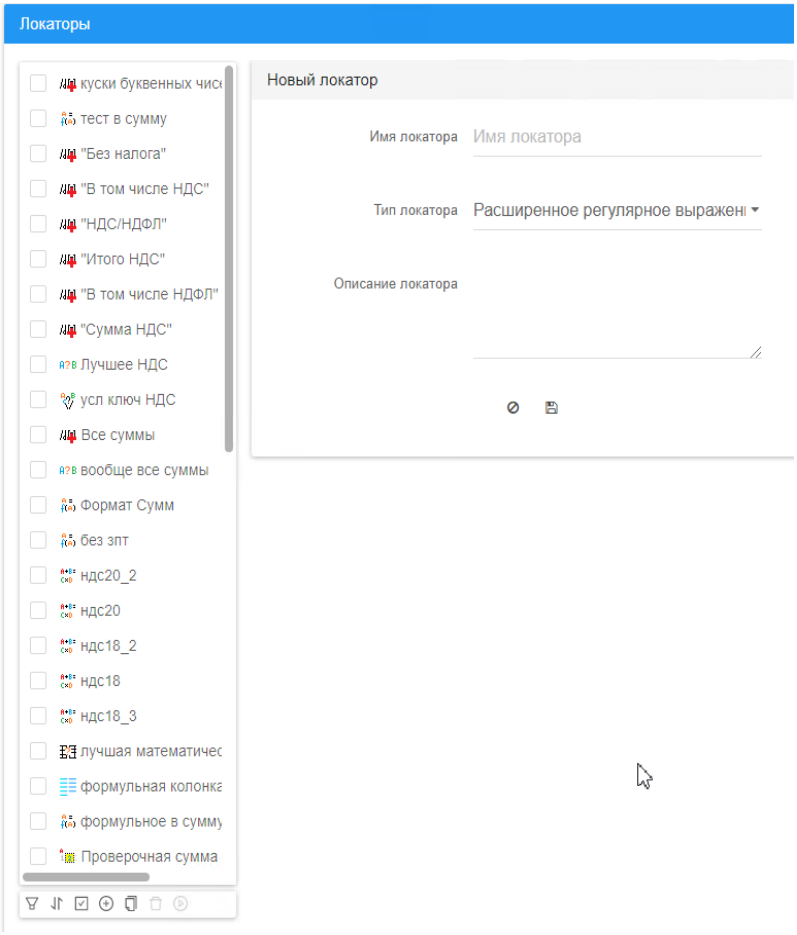
- Поле содержит в имени **«singlecompare»**. В тексте данного поля должно содержаться правило проверки полей и таблиц внутри документа следующего вида: **Имя_класса1.Имя_поля1/Имя_таблицы1=Имя_класса1.Имя_поля1/Имя_таблицы1.** Для полей все как в предыдущем случае. В таблицах – становятся невалидными несовпадающие ячейки.
- Что бы сравнить лица на документах разных классов внутри пакета нужно создать поле пакета содержащее в имени **«facecompare»**, в тексте которого должна быть конструкция типа: **{ИмяКлассаДокумента1}.{ИмяПоля1}={ИмяКлассаДокумента2}.{ИмяПоля2}. Например: Паспорт.Фото=Права.Фото.**

Если лица в полях совпадают меньше чем на 50%, то поля станут не валидными с сообщением: "Лицо в поле Фото документа класса Паспорт совпадает с лицом в поле Фото документа класса Права лишь на N%".

2.4 Меню поиска данных (Классы документов).

В меню поиска данных настраиваются инструменты для поиска данных с изображения и вид, в котором данные будут переданы на валидацию. Все настройки в этом меню относятся к отдельному выбранному классу документа.

2.4.1 Локаторы.



Локаторы — это инструмент нахождения (извлечения) данных с выбранного документа. Каждый локатор имеет свои настройки.

Локаторы могут быть связаны между собой и выполняются последовательно сверху вниз.

Результаты работы разных локаторов могут сравниваться между собой для выявления лучшего итогового результата.

Результаты работы локатора могут выступать ключами для других локаторов (т.е. отработка других локаторов будет производиться относительно результатов текущего).

Всего в системе 37 локаторов.

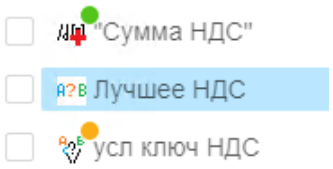
Каждый локатор имеет свой значок, для удобство зрительной идентификации.

(Рис. 58 Локаторы. Общий вид)

У некоторых локаторов есть несколько подполей к которым можно обращаться впоследвии. Например: ключ поиска, регион поиска и результат поиска. Так же у результата поиска может быть несолько альтернатив. Это значит по указанным условиям поиска нашлось несколько совпаденений на изображении.

У каждого результата OCR есть своя степень доверия. Это % с которым движок распознования уверен что найденый результат соответвует данным с изображения. Степень доверия играет большую роль в работе с локаторами. Каждому результату локатора передается степень доверия от рузультата OCR.

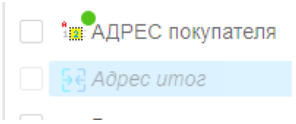
В одном классе документа локаторы связаны друг с другом. Для удобства связанные локаторы имеют цветовой индикатор.



Зеленый – локатор от которого зависит выбранный (наследуемый).

Оранжевый – локатор, который зависит от выбранного (зависимый).

У выбранного локатора можеть быть много наследуемых и зависимых локаторов.



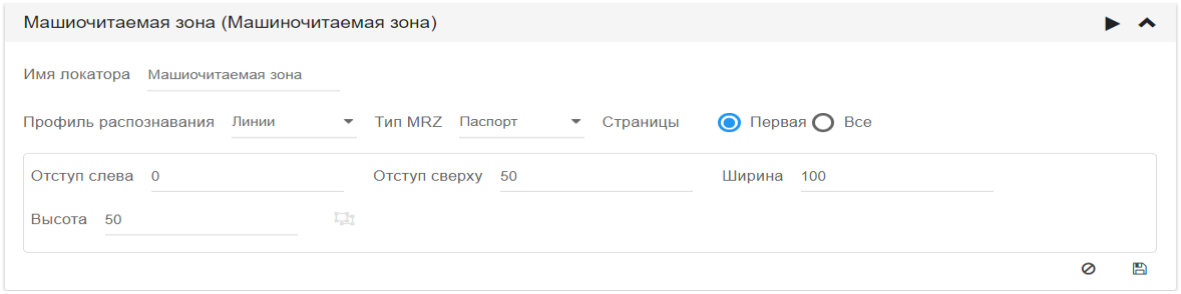
Локатор, не имеющий зависимого локатора или не привязанный к полю (таблице), выделен серым курсивом.

1. Машиночитаемая зона (MRZ)

Локатор, осуществляющий поиск машиночитаемой зоны с документов удостоверяющих личность, таких как паспорта. Машиночитаемая зона – это основная информация с документа, зашифрованная определенным набором символов.

(Рис. 60 Пример mrz зоны)

Тип извлекаемой зоны – паспорт (2 строки). Строки ищутся на указанном профиле: необходимо чтобы профиль включал английский язык; лучшим решением будет оставить только английский язык в профиле, также установить ограничение на список находимых символов – большие латинские буквы, арабские цифры и знак “<”. Этот знак (меньше) является одним из маркеров нахождения машиночитаемой зоны. Результатом работы локатора является набор подполей – частей mrz строк, как с данными, так и с контрольными суммами для валидации данных. Количество альтернатив в каждом подполе 0 или 1. Осуществляется поиск только на первой странице документа.



(Рис. 59 Локатор Машиночитаемая зона)

Профиль распознавания. Указывает репрезентацию (результаты профиля распознавания) в которых будет осуществлен поиск строк машиночитаемой зоны.

Тип MRZ (машиночитаемой зоны). Указывает алгоритм поиска и разбора строк машиночитаемой зоны. Возможные значения – passport. В этом случае машиночитаемая зона состоит из 2х строк и разбирается по следующему принципу: https://en.wikipedia.org/wiki/Machine-readable_passport

Страницы, на которых будет выполнятся инструмент извлечения. Возможные варианты – первая страница. При поиске на **первой странице**, будет осуществляться поиск MRZ строк в результатах ocr указанного профиля только первой страницы документа.

Настройки области инструмента извлечения. Описывает область репрезентации, в которой будет выполняться поиск mrz. Область можно задавать графически на репрезентации.

- **Отступ слева** – отступ от левой границы изображения репрезентации выраженный в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.
- **Отступ сверху** – отступ от верхней границы изображения репрезентации выраженный в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.
- **Ширина** – ширина области, выраженная в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.
- **Высота** – высота области, выраженная в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.

Суммы Отступ слева, Ширина и Отступ сверху, Высота так же не могут быть больше 100.

Подполя MRZ:

"Тип" – тип документа

"Страна" – код государства, выдавшего документ

"Фамилия" – фамилия владельца документа

"**Имя**" – имя (имена) владельца документа

"**Номер**" – номер документа

"**Контрольная сумма 1-9**" – контрольная сумма символов 1-9

"**Национальность**" – национальность владельца документа

"**Дата рождения**" – дата рождения владельца документа

"**Контрольная сумма 14-19**" – контрольная сумма 14-19

"**Пол**" – пол владельца документа

"**Дата истечения срока действия**" – дата истечения срока действия документа

"**Контрольная сумма 22-27**" – контрольная сумма 22-27

"**Личный номер**" – личный номер владельца документа

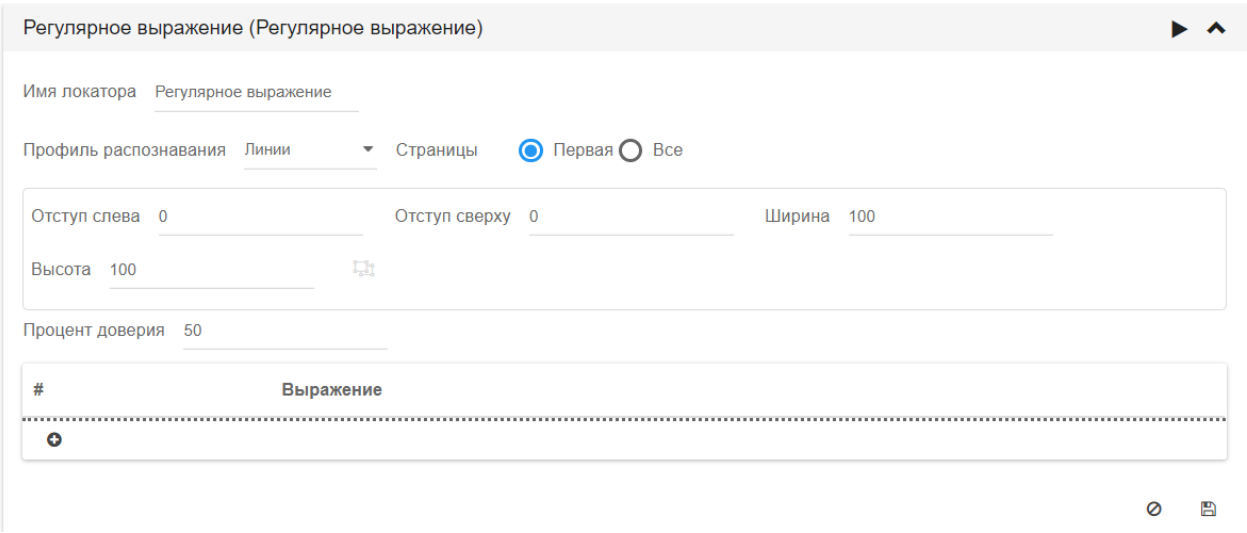
"**Контрольная сумма 29-42**" – контрольная сумма 29-42

"**Контрольная сумма**" – контрольная сумма контрольных сумм

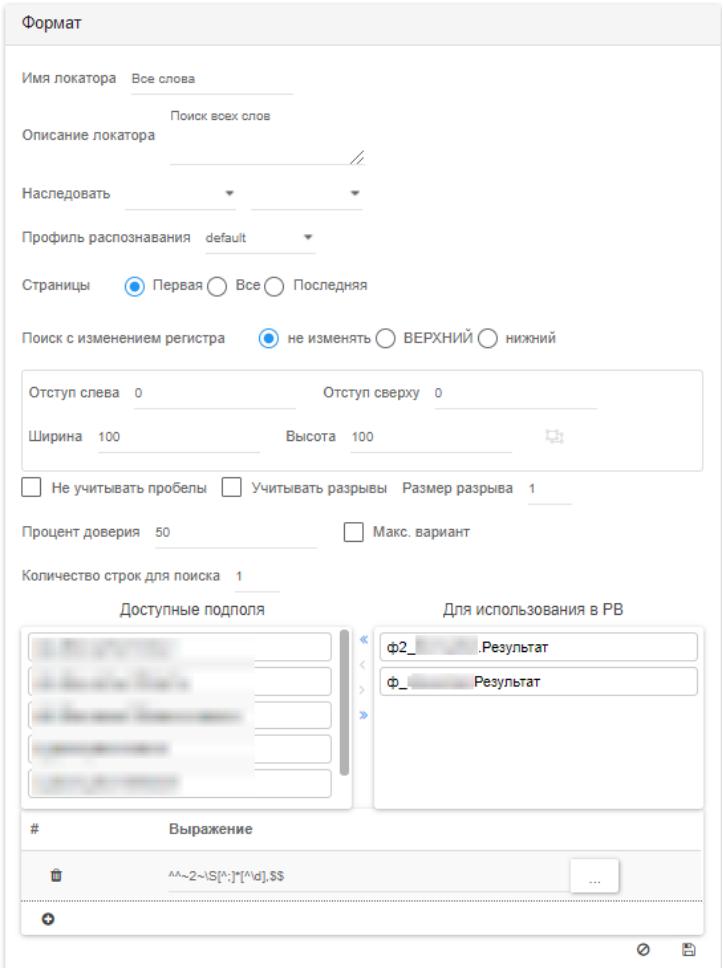
Максимальное количество альтернатив для каждого поля – 1.

2.Регулярное выражение ( ) и Расширенное регулярное выражение ( )

Инструмент извлечения осуществляет поиск слов, соответствующих одному из регулярных выражений в указанной области результата распознавания документа. Например регулярное выражение - а(м|ш)а , может вернуть слова Мама, Маша, ашан и т.д. Эти слова будут альтернативами единственного подполя инструмента извлечения. В расширенном регулярном выражении присутствует больше настроек.

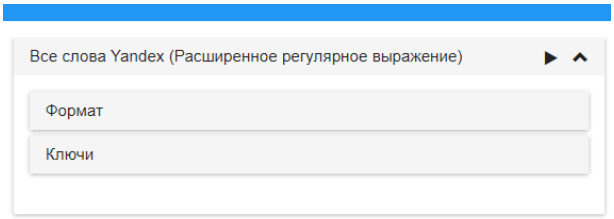


(Рис. 62. Локатор «Регулярное выражение»)



(Рис. 63 Локатор «Расширенное регулярное выражение)

Локатор «Расширенное регулярное выражение» состоит из двух разделов настроек, «Формат» и «Ключи».



Раздела «Формат» содержит следующие настройки (Рис. 63 Локатор «Расширенное регулярное выражение):

Наследование подполя инструмента извлечения. Альтернативы наследуемого подполя используются как регион поиска регулярных выражений.

Профиль распознавания. Указывает репрезентацию (результаты профиля распознавания) в которых будет осуществлен поиск указанных регулярных выражений.

Настройки области инструмента извлечения. Описывает область репрезентации, в которой будет выполняться поиск указанных регулярных выражений. Область можно выбирать графически (с помощью выделения прямоугольника на документе), либо задать область числами (в процентах от размера страницы).

- **Отступ слева** – отступ от левой границы изображения репрезентации выраженный в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.
- **Отступ сверху** – отступ от верхней границы изображения репрезентации выраженный в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.
- **Ширина** – ширина области, выраженная в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.
- **Высота** – высота области, выраженная в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.

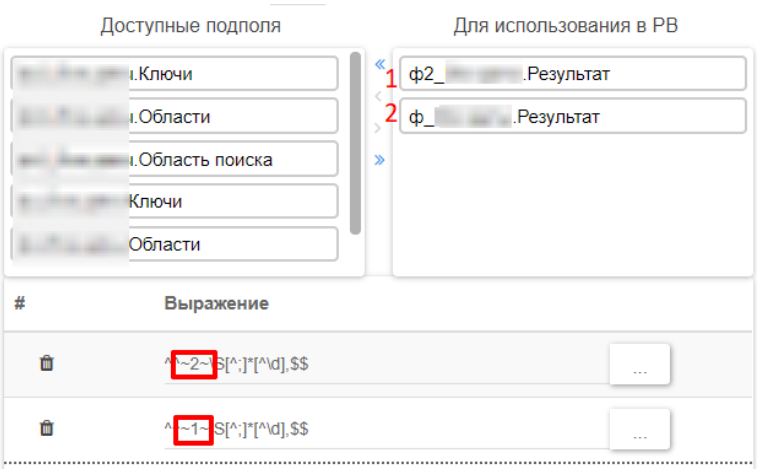
Суммы Отступ слева, Ширина и Отступ сверху, Высота так же не могут быть больше 100.

Страницы, на которых будет выполнятся инструмент извлечения. Возможные варианты – первая страница, все страницы, последняя. Поиск результатов ocr будет осуществляться в соответствии с выбранными значением. Если выбрана первая (либо последняя) страница, то поиск результатов не будет производиться на остальных страницах.

Минимальный процент доверия. Ограничивает результаты выбора слов из репрезентации по проценту доверия (учитывается коррекция по ключам). Диапазон значений: 0-100. Чем выше процент, тем сильнее будет отбор результатов ocr с выбранной репрезентации. Процент доверия вероятной альтернативы должен быть выше чем указанный порог в локаторе.

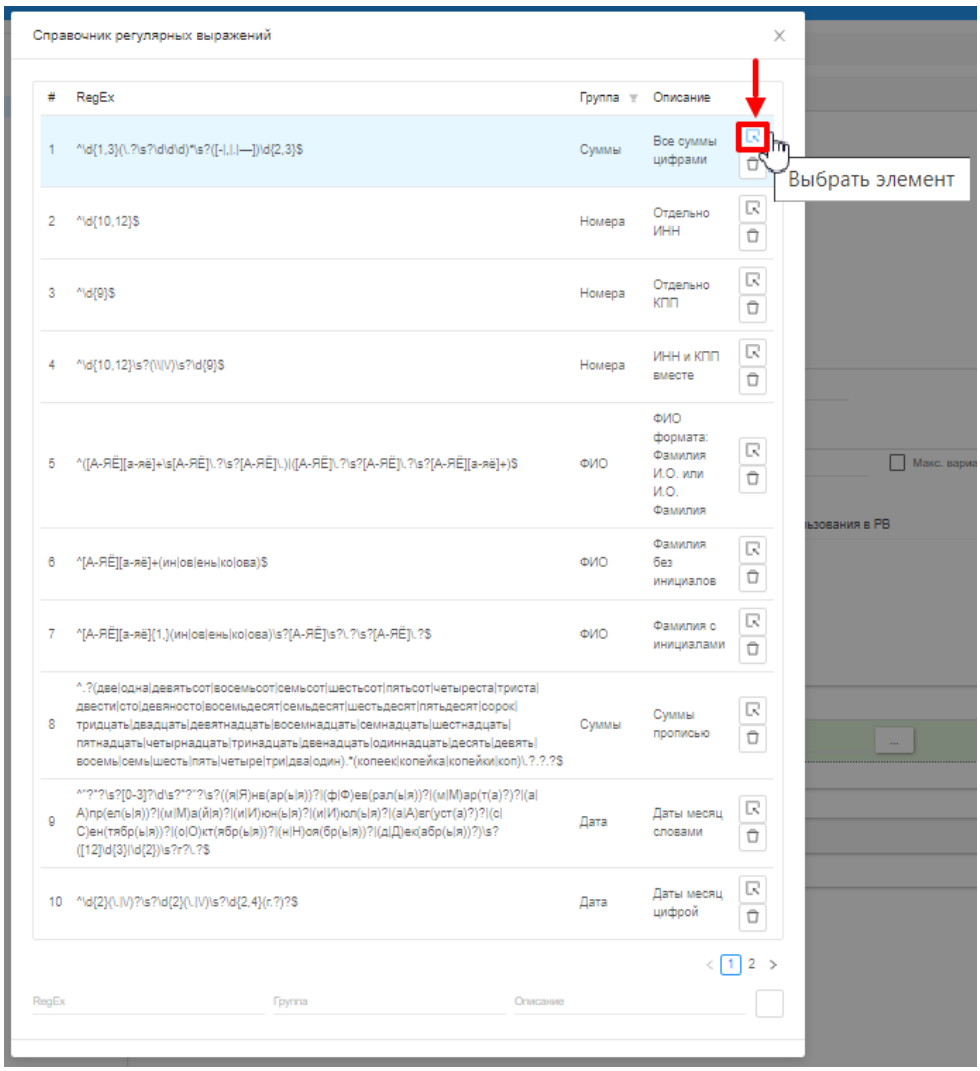
В **доступных полях** отображаются альтернативы всех локаторов предшествующих локатору расширенного регулярного выражения. Любую альтернативу можно использовать внутри регулярного выражения. Для этого необходимо перенести нужную альтернативу в список **Для использования в РВ.** А при написании регулярного выражения внутри символов «~» написать порядковый номер альтернативы из получившегося списка.

Пример:

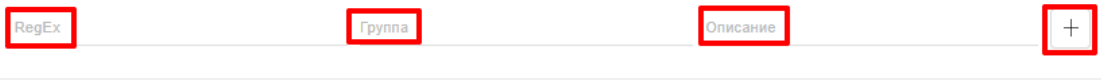


Список регулярных выражений для поиска. В нем указываются регулярные выражения с корректным синтаксисом, на которые будут проверяться слова и совокупности слов из указанной области репрезентации. При соответствии слова выражению, слово будет записано в альтернативу подполя инструмента извлечения. В регулярное выражение можно включить значения из справочника в формате **§-source-имя_справочника§**. Для того чтобы указать что символ должен быть именно началом строки, либо концом строки (то есть до него, либо после него нет никаких символов в строке) необходимо поставить двойное начало строки, либо двойной конец строки **(^^ или $$)**

С помощью кнопки **Справочника регулярных выражений** можно выбирать регулярные выражения из имеющихся в базе. Чтобы добавить регулярное выражение из справочника необходимо нажать на соответствующую кнопку или щёлкнуть двойным кликом по нужному выражению из списка.



Для добавления в базу регулярного выражения необходимо ввести само выражение RegEx, название группы Группа и краткое описание Описание. После этого нажать на кнопку добавить +.



Не учитывать пробелы. Данная настройка позволяет игнорировать пробельные символы в сочетании слов при проверке на совпадение с регулярным выражением.

Учитывать разрывы. Данная настройки позволяет запрещать выполнять проверку на регулярное выражение сочетания слов, если расстояние по горизонтали, больше чем указанное значение.

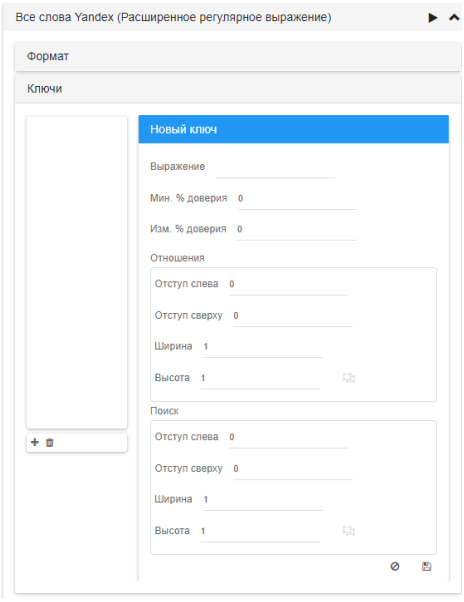
Размер допустимого разрыва. Указывает величину допустимого для объединения слов разрыва в количестве усредненных по ширине символов соседних слов. Условно указывается допустимое расстояние между словами, равное количеству символов, которые могут влезть между словами.

Максимальный вариант. Данная настройка позволяет выбирать из пересекающихся по координатам альтернатив результирующего подполя инструмента извлечения, только одну альтернативу длина текста у которой наибольшая, а остальные альтернативы удаляются. Из всех найденных альтернатив берется максимально длинная альтернатива, если она пересекается с остальными.

Количество строк для поиска. Количество строк подряд, в которых будет выполнен поиск регулярных выражений. При указывании количества строчек для поиска 0. Выполняется запись всей строки независимо от содержимого в альтернативу локатора.

Регистр текста. Варианты использования: ВЕРХНИЙ, нижний, не изменять. Данная настройка позволяет переводить результат OCR в подконтрольный регистр для более гибкого использования условий регулярных выражений и поиска значений в базах данных. К указанному регистру приводится искомый текст перед сравнением с регулярным выражением.

В настройках локатора может присутствовать **список ключей**. Ключи служат для изменения степени доверия альтернатив. Каждый ключ содержит Текст ключа (Выражение), минимальный процент доверия ключа, процент изменения доверия ключом, настройки области поиска ключа, настройки расчета области применения ключа.



(Рис. 63.1 Раздел настроек «Ключи»)

Список ключей настраивается в разделе «Ключи», который содержит следующие настройки:

Текст ключа. Текст, с которым сравниваются слова из ocr с указанной в локаторе репрезентации. Сравнение выполняется по неточному совпадению, основанному на расстоянии Левенштайна http://en.wikipedia.org/wiki/Levenshtein_distance

Минимальный процент доверия ключа. Процент доверия ключа рассчитывается как произведение процента доверия слова в ocr и процента совпадения слов деленное на 100. Если этот процент меньше указанного, то ключ не участвует в дальнейших расчетах. Диапазон значения настройки 0-100.

Процент изменения доверия ключом. Настройка, указывающая на сколько будет изменен процент доверия альтернативы попавшей в область применения ключа. При попадании в область ключа процент доверия альтернативы пересчитывается следующим образом: текущий процент доверия + (произведение процента доверия слова ключа в ocr и процента совпадения слов ключа и слова в ocr)/100*процент изменения доверия ключом/100. Диапазон значения настройки -100-100.

Настройки области поиска ключа (аналогичны настройкам области инструмента извлечения). Описывает область репрезентации, в которой будет выполняться поиск ключа.

Настройки расчета области применения ключа. Описывают область на репрезентации инструмента извлечения, степень доверия попавших альтернатив инструмента извлечения полностью в которую будет изменяться в соответствии с указанными значения для ключа.

Если альтернатива попадает в несколько областей ключей, то к ней будет применено несколько правил. Каждый ключ может быть найден на документе несколько раз.

Подполя инструмента извлечения:

Результат – слова или совокупности слов из указанной области ocr указанной репрезентации первой или всех страниц документа, у которых процент доверия выше указанного в инструменте извлечения, и которые соответствуют хотя бы одному из указанных регулярных выражений, либо альтернативы унаследованные от указанного в настройках подполя инструмента извлечения, выполняющегося ранее.

Ключи – слова из указанных областей ocr указанной репрезентации первой или всех страниц документа, у которых произведение процента доверия в ocr и процента совпадения с одним из ключей деленное на 100 больше или равно указанному минимальному значению степени доверия ключа.

Области – области ключа для найденных ключей, рассчитанных исходя из размера и расположения ключа и настроек расчета области применения ключа.

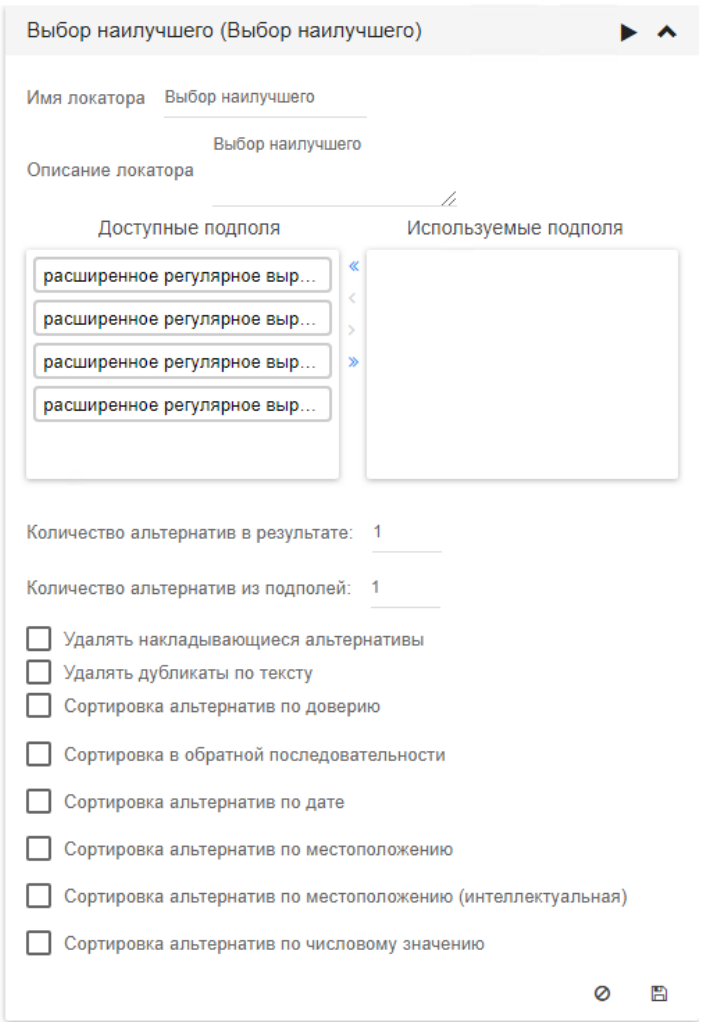
Для каждого ключа(слова) необходимо создать свои настройки (новый ключ). В одном ключе может использоваться только одно слово.

3. Выбор наилучшего (A?B).

Инструмент извлечения позволяет выбирать заданное количество альтернатив из других локаторов в соответствии с правилами.

указанное количество лучших по степени доверия альтернатив указанных подполей инструментов извлечения.

Подполе у инструмента извлечения одно.



(Рис. 64 Локатор «Выбор наилучшего»)

Если ни один пункт не выбран, то альтернативы будут отсортированы по мере вхождения в локатор выбора наилучшего. Т. е. первая альтернатива первого локатора в списке будет на первом месте, а последняя альтернатива последнего локатора в списке будет на последнем.

Используемые подполя. Список подполей инструментов извлечения, из альтернатив которых будут выбираться альтернативы для текущего инструмента извлечения.

Количество альтернатив из подполей. Максимальное количество лучших (с наибольшей степенью доверия) альтернатив наследуемых подполей инструментов извлечения, которые участвуют в дальнейшем выборе. Диапазон значений: 1-100.

Количество альтернатив в результате. Максимальное количество выбранных альтернатив из общего списка наследуемых альтернатив, отсортированного по степени доверия в порядке ее уменьшения. Диапазон значений: 1-100.

Удалять пересекающиеся альтернативы. Опция, позволяющая удалять в результирующем подполе локатора альтернативы накладывающиеся геометрически на другие альтернативы.

Удалять дубликаты. Опция, позволяющая удалять в результирующем подполе локатора альтернативы имеющиеся тот же текст что и другие альтернативы.

Сортировать по доверию. Альтернативы всех локаторов сортируются в зависимости от степени доверия ocr и берется только указанное количество альтернатив. Например, у локаторов л1 и л2, есть подполя п1 и п2 соответственно. Альтернативы л1.п1: 1-95%, 2-60%, 3-12%; л2.п2: 1-73%, 2-44%. Если в настройках указано что берем по 2 лучших альтернатив из инструментов извлечения и в результате будет 3 лучших, то в результате будут альтернативы: л1.п1.1 – 95%, л2.п2.1 – 73%, л1.п1.2 – 60%.

Сортировать в обратной последовательности. Альтернативы будут отсортированы в обратном порядке вхождения альтернатив в локатор выбор наилучшего. Т. е. первая альтернатива первого локатора в списке будет последней, а последняя альтернатива последнего локатора будет первой.

Сортировка альтернатив по дате. Даты, попавшие в локатор, будут отсортированы от меньшего к большему. Т. е. свежие даты будут первыми, а более ранние даты в конце (01.01.2000,31.12.1999,01.01.1998 и т. д.)

Сортировать по местоположению. Опция, которая позволяет выполнять сортировку альтернатив в результирующем подполе локатора исходя из их координат – слева направо и сверху вниз.

Сортировать по местоположению (интелектуальная).

Сортировка альтернатив по числовому значению. Альтернативы будут отсортированы от большего значения к меньшему. Т. е. на первом месте будет максимальное числовое значение. Если комбинировать данную сортировку с функцией обратной сортировки, то на первом месте будет меньшее значение.

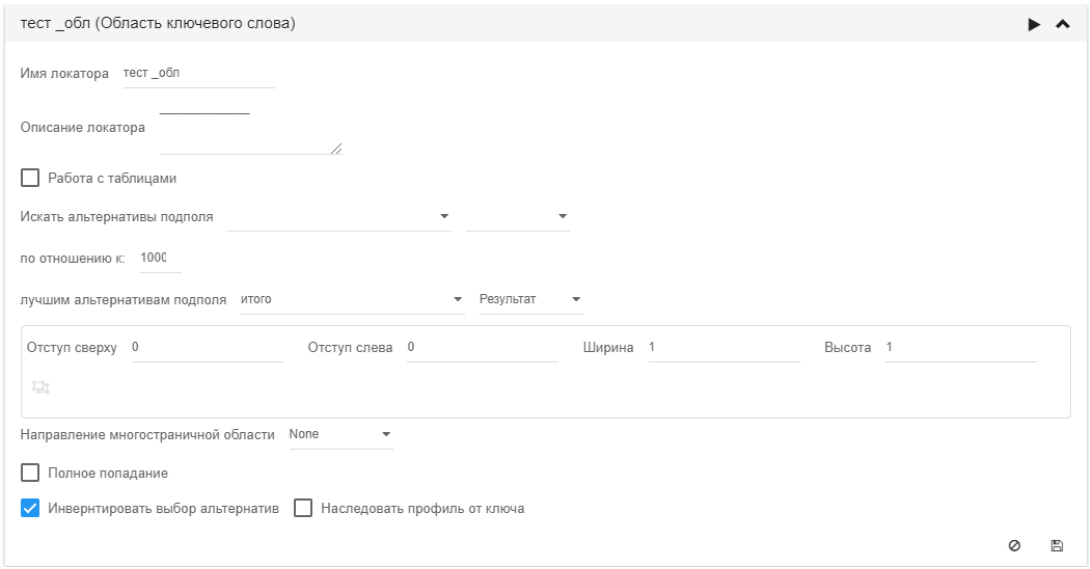
Подполя инструмента извлечения:

Результат. Содержит отсортированный перечень альтернатив в соответствии с указанными настройками.

4. Область ключевого слова ( )

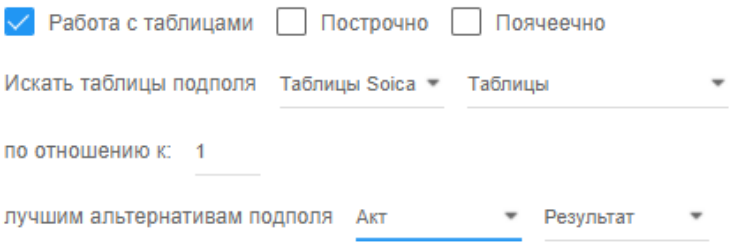
Позволяет отсеивать таблицы (можно отсеивать строки таблиц) или альтернативы одного подполя локатора относительно расположения альтернатив другого подполя локатора. Расположение определяется областью, координаты и размеры которой выражены в долях от ключевой альтернативы подполя. Ключевая альтернатива может быть одна или несколько. Например, нам нужны номера, которые находятся справа от ключевых слов «накладная» и «договор», ищем ключевые слова предыдущими локаторами, ищем все номера предыдущими локаторами, и локатором область ключевого слова отсеиваем вторые относительно первых. Номера, которые не попали в области ключей, (допустим, шириной 2 ключа, высотой 3 ключа, расположенной справа от ключа) не записываются в результат конечного локатора.

Условно локатор области ключевого слова ищет уже найденные альтернативы в определенной области относительно указанного ключа. В качестве ключа так же берется заранее найденный результат локатора.



(Рис. 65 Локатор «Область ключевого слова»)

Работа с таблицами. При выбранной опции будут отсеиваться таблицы, строки или ячейки таблиц. Для того, чтобы отсеивались строки таблиц необходимо дополнительно выбрать опцию «Построчно». Для того, чтобы отсеивались ячейки таблиц необходимо дополнительно выбрать опцию «Поячеечно».



Искать альтернативы подполя (Наследуемое подполе). Подполе локатора, альтернативы которого записываются в результирующее подполе, при попадании их в указанные ниже области.

Лучшим альтернативами подполя (Ключевое подполе). Подполе локатора, альтернативы которого образуют ключи, относительно которых рассчитываются ключевые области.

По отношению к: (Количество ключевых альтернатив). Указывает количество лучших (по степени доверия) альтернатив ключевого локатора, которые образуют ключи. Диапазон значений: 1-100.

Настройки расчета области применения ключа. Описывают область, при полном попадании в которую альтернативы наследуемого подполя будут добавляться в результирующее подполе текущего локатора. Области можно задавать строго числами, либо графически на репрезентации.

- **Отступ слева** – отступ от левой границы ключа выраженный в долях от ширины ключа. Диапазон значений: -25.00-25.00.
- **Отступ сверху** – отступ от верхней границы ключа выраженный в долях от высоты ключа. Диапазон значений: -25.00-25.00.

- **Ширина** – ширина области, выраженная в процентах от ширины ключа. Диапазон значений: 0.00-25.00.
- **Высота** – высота области, выраженная в процентах от высоты ключа. Диапазон значений: 0.00-25.00.

Полное попадание в регион. Опция, указывающая, должны ли альтернативы полностью геометрически попадать в область ключа, или достаточно лишь частичного наложения.

Инвертировать выбор альтернатив. Позволяет не «выбирать те, что попали в область», а наоборот «те что не попали в область».

Наследовать профиль от ключа. Позволяет поменять репрезентацию альтернатив на репрезентацию ключей

Направление многостраничной области. Позволяет переносить область поиска на другие страницы по вертикали или горизонтали

Подполя инструмента извлечения:

Результат. Альтернативы наследуемого подполя, которые полностью попали в одну или несколько областей рассчитанных, из указанного количества альтернатив ключевого подполя по их координатам и настройкам расчета области применения ключа.

Ключи – указанное количество лучших (по степени доверия) альтернатив ключевого подполя локатора.

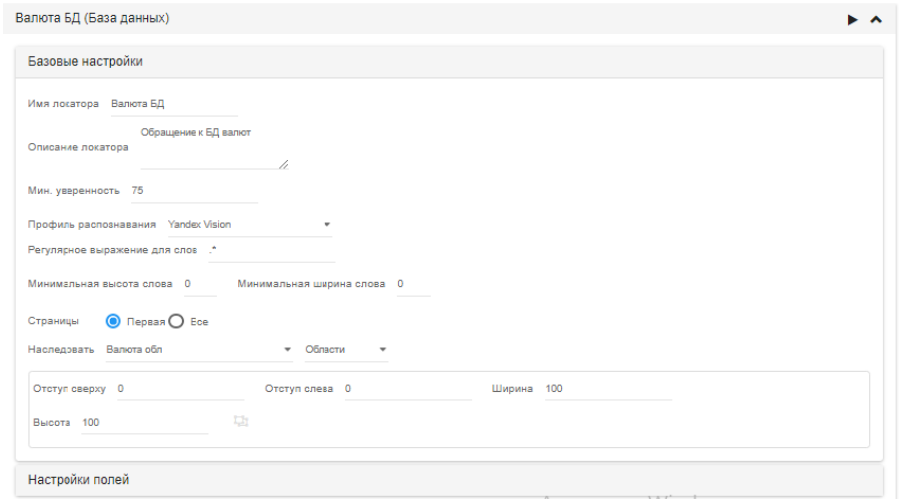
Области – области ключа для найденных ключей, рассчитанных исходя из размера и расположения ключа и настроек расчета области применения ключа.

5. База данных ( )

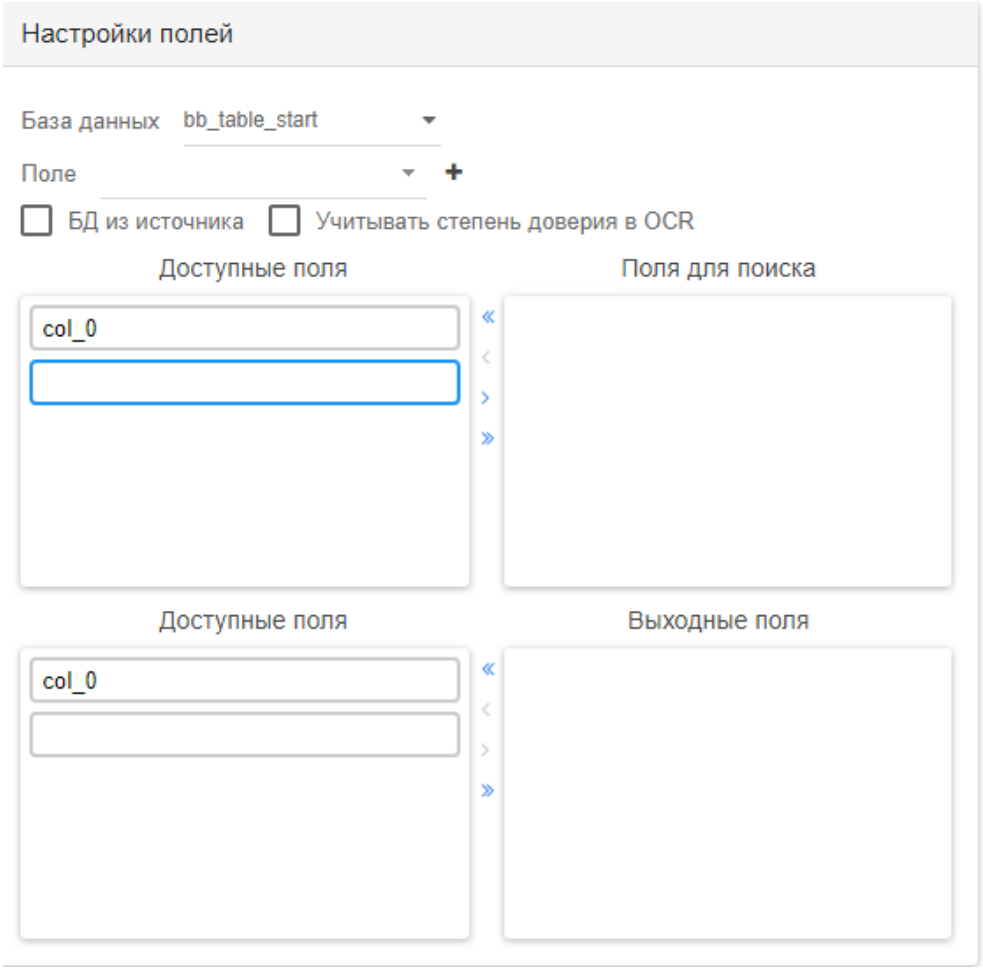
Локатор осуществляет поиск в определенной области на указанной репрезентации документа слов или совокупностей слов, наиболее соответствующих одной из строк из столбцов выбранного источника данных. Затем происходит вычисление наиболее подходящих строк из источника данных по указанным полям. И в порядке убывания степени доверия строк из источника данных в подполя локатора записываются указанные выходные поля. Результаты так же отсеиваются минимальной степенью доверия с записи из источника данных.

Допустим нам необходимо найти данные по поставщику на странице, имея базу данных поставщиков. Мы знаем, что на странице должно быть наименование и ИНН поставщика, а нам нужен ОГРН. Необходимо выбирать базу поставщиков в качестве источника, выбирать поля для поиска: Наименование и ИНН. Затем указать поля в результат: ОРГН и ограничить область поиска. Так же часто целесообразно указать высокую минимальную степень доверия к результатам поиска, для того чтобы если поставщика нет в базе, то имеющиеся записи в базе не пытались добавится в результат поиска. При степени доверия 0, все записи из базы попадут в результат в порядке уменьшения степени доверия.

Помимо статично указанной области (в % от размера страницы) есть возможность наследовать область из альтернативы подполя другого локатора.



(Рис. 66 Локатор «База данных»)



Профиль распознавания. Выбор репрезентации (результаты профиля распознавания) в которой будет осуществлен поиск значений из указанных полей источника данных.

Регулярное выражение для слова. Позволяет выбрать те слова результатов распознавания, которые будут участвовать в поиске.

Минимальная ширина слова, минимальная высота слова. Данные настройки позволяют задать параметры для высоты и ширины слова, результаты меньше которых не

попадут в альтернативы локатора. Высота и ширина измеряются в десятых долях процента от размера изображения страницы.

Настройки «Регулярное выражение для слова», «Минимальная ширина слова», «Минимальная высота слова» сделаны для увеличения скорости работа локатора, чтобы не искать в мусоре.

Настройки области инструмента извлечения. Описывает область репрезентации, в которой будет выполняться поиск значений из указанных полей выбранного источника данных. Область можно задавать графически на реперзинации.

- **Отступ слева** – отступ от левой границы изображения репрезентации выраженный в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.
- **Отступ сверху** – отступ от верхней границы изображения репрезентации выраженный в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.
- **Ширина** – ширина области, выраженная в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.
- **Высота** – высота области, выраженная в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.
- Суммы Отступ слева, Ширина и Отступ сверху, Высота так же не могут быть больше 100.

Страницы, на которых будет выполнятся инструмент извлечения. Возможные варианты – первая страница, все. При поиске на **первой странице**, будет осуществляться поиск значений из указанных полей источника данных в результатах ocr профиля только первой страницы документа. При поиске на **всех страницах**, будет осуществляться поиск значений из указанных полей источника данных в результатах ocr профиля на каждой из страниц документа.

Мин уверенность. Ограничивает результаты выбора строк из источника данных по рассчитанному проценту доверия как среднеарифметическое по каждому указанному полю из источника данных: из процента доверия сравниваемого слова или совокупности слов и процента совпадения слова или совокупности слов со значением поля в источнике. Диапазон значений: 0-100.

Наследуемое подполе области. Если указано это подполе, то регион уже не рассчитывается исходя из настроек области инструмента извлечения. В качестве региона берутся координаты и размер лучшей (по степени доверия) альтернативы указанного подполя локатора. Если область наследуемой альтернативы выходит за границы изображения репрезентации, то итоговый регион локатора корректируется. Если альтернатив у наследуемого подполя нет, то локатор не выдаст результатов.

База данных. Ссылка на таблицу с данными, ячейки строк которой будут сопоставляться с результатами ocr указанной репрезентации.

БД из источника. При включенной опции, в момент выполнения локатора, таблица с данными будет наполняться из файла или подключенной таблицы из базы данных. Иначе таблица будет наполняться из временной копии, сохраненной в базе данных системы.

Поле. Список полей из таблицы источника данных, ячейки из которых будут сопоставляться со словами или совокупностями слов в указанном регионе результатов ocr репрезентации.

Выходные поля и итоговое поле. Список полей источника данных, которые будут сопоставляться с выходными подполями локатора.

Учитывать степень доверия OCR. Опция, указывающая, будет ли степень доверия слов в OCR влиять на расчет процента доверия альтернатив локатора.

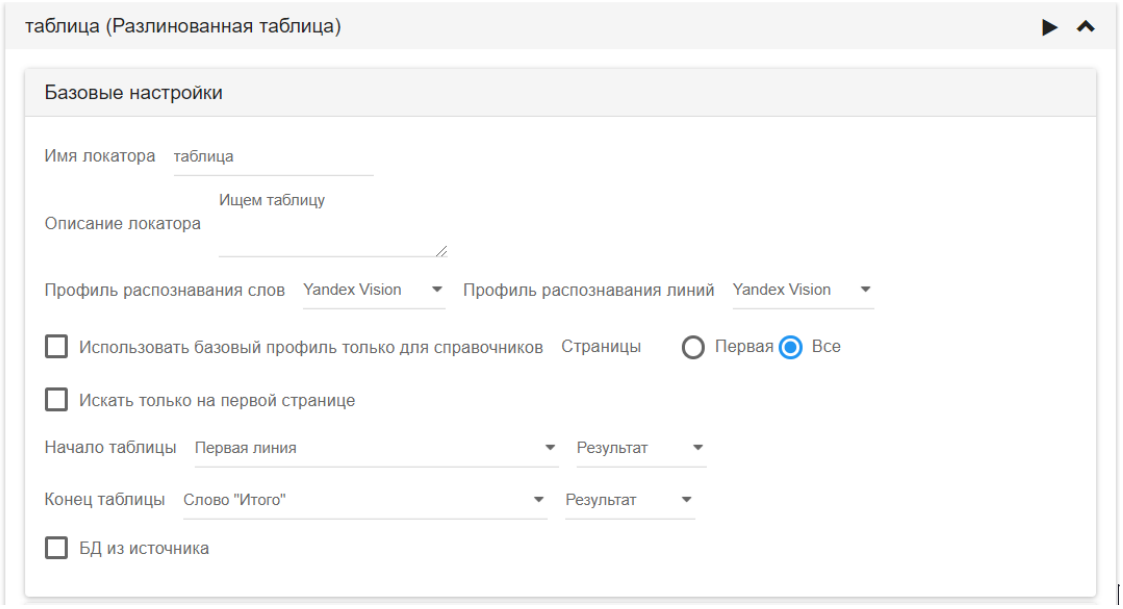
Подполя инструмента извлечения:

Выходные подполя. Содержат альтернативы, заполненные из ячеек таблицы источника данных, отсортированные в порядке убывания степени доверия итоговой строки из источника данных.

6. Разлинованная таблица ( )

Локатор служит для поиска табличных данных. Сначала ищутся линии и образуются сетки, а затем в них находятся ocr указанного профиля распознавания. Локатор удобно использовать на документах с четко выделенными линиями в таблице, в противном случае необходимо будет выполнять тонкую настройку поиска линий для таблицы на документе. Профиль для поиска линий и поиска данных в таблице может быть разным.

Подполя: таблица, начало, конец, левая и правая границы.



(Рис. 67 Локатор «Разлинованная таблица»-Базовые настройки)

Профиль распознавания слов. Указывает репрезентацию, из результатов ocr которой будут извлекаться слова для наполнения ячеек таблицы.

Профиль распознавания линий. Указывает репрезентацию, в которой будет осуществлен поиск линий.

Использовать базовый профиль только для справочников. Опция, указывающая будет ли профиль для слов использоваться для заполнения содержимого таблицы или только лишь для поиска начала и конца таблицы.

Страницы, на которых будет выполнятся инструмент извлечения. Возможные варианты – первая страница, все страницы. При поиске на **первой странице**, будет осуществляться поиск одностраничной таблицы только в указанных реперезентациях первой страницы документа. При поиске на **всех страницах**, будет осуществляться поиск многостраничной таблицы только в указанных реперезентациях всех страниц документа.

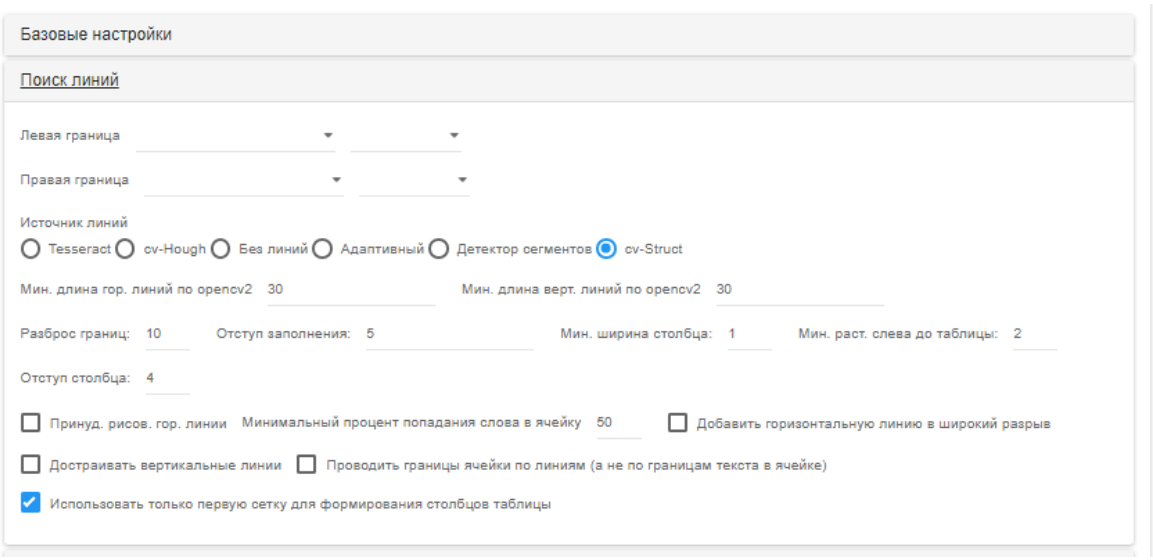
Начало таблицы. Подполе заранее выполненного локатора или БД для нахождения начала таблицы. Может быть не указан.

Искать на первой странице документа. В случае, если эта опция выбрана, поиск начала таблицы по справочнику будет осуществляться в результатах ocr указанной репрезентации для поиска слов только на первой странице документа.

Конец таблицы. Подполе локатора или БД для нахождения конца таблицы. Может быть не указан.

БД из источника. При выборе данного пункта необходимо указать таблицу для нахождения начала и конца таблицы, значения строк которой будет сравниваться со строками в выбранной репрезентации для поиска слов, указанных страниц. В качестве начала таблицы выбирается строка результатов ocr которая наиболее совпадает со строкой в ocr, в случае если процент совпадения строк выше 80.

Заполнение таблиц осуществляется по результатам ocr с указанного профиля.



(Рис. 67.1 Локатор «Разлинованная таблица»-Поиск линий)

Левая граница таблицы. Указывает подполе, лучшая (по степени доверия) альтернатива которого будет являться левой границей таблицы.

Правая граница таблицы. Указывает подполе, лучшая (по степени доверия) альтернатива которого будет являться правой границей таблицы.

«**Добавить горизонтальную линию в широкий разрыв**» - в случае, когда между строками ocr в таблице, расстояние больше чем средняя высота строки, то добавляется линия таблицы, если эта опция выбрана. Ранее активировалось костылем «vg»

«**Минимальный процент попадания слова в ячейку**» - процент площади слова, которая должна пересекаться с площадью ячейки, для того чтобы оно считалось

принадлежавщей ей. По умолчанию – 0 (при любом пересечении), костыль «fill50» - менял этот процент на 50.

«**Использовать только первую сетку для формирования столбцов таблицы**» - по умолчанию используется только первая сетка для формирования таблицы, если использовался костыль «all_grids», то сетки использовались все.

«**Проводить границы ячейки по линиям (а не по границам текста в ячейке)**» - если не использовался костыль «full_cells», то границы ячейки определялись границами слов в ней, иначе – линиями таблицы.

Источник линий. Позволяет указать алгоритм, которым будет выполняться поиск линий, для построения сеток таблицы. Варианты: Tesseract, cv-Hough, без линий, Адаптивный, Детектор сегметов, cv-Struct

Алгоритмы поиска линий: 1 – основанный на поиске границ методом Кэнни и преобразование их в линии методом Хафа; 2 – основанный на морфологических преобразованиях; 3 – использование линий, найденных Tesseractom; 4 – построение линий исходя из структуры результатов распознавания; 5 – построение линий исходя из графической структуры документа.

Первый способ обладает рядом настроек и менее универсальный.

Второй способ предпочтителен для таблиц, имеющих границы. Так же, как и первый не требует результаты ocr для работы. В большинстве случаев качественнее и быстрее первого.

Третий способ подходит для документов хорошего качества, но требует результаты распознавания. При наличии последних, является самым быстрым.

Четвертый способ, как и предыдущий требует результаты ocr. Служит для расчета линий, когда таблица не разлинована. Выполняется анализ структуры результатов распознавания. Обладает широким рядом настроек. Подходит для узкого круга задач.

Пятый способ не требует результатов ocr. В остальном имеет схожую логику что и четвертый, настроек нет.

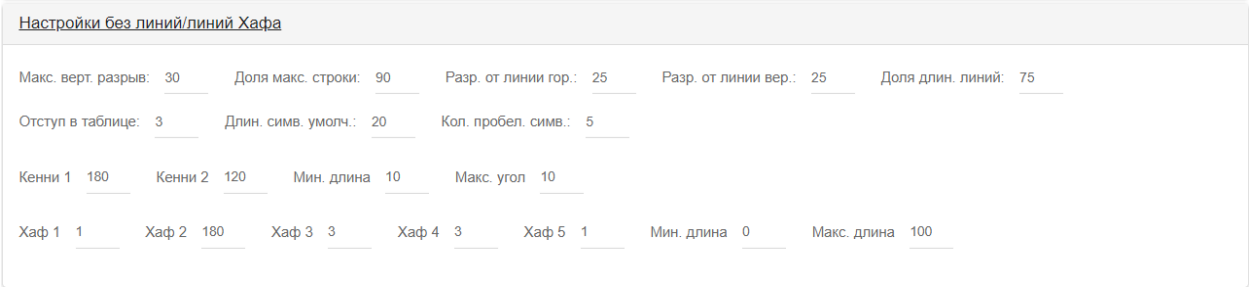
Поиск линий выполняется на указанной репрезентации.

Далее найденные линии объединяются в сетки, которые объединяются в таблицы, в том числе многостраничные.

Для 4го и 5го способа поиска линий поиск начала и конца таблицы зачастую является обязательным условием. Поиск начала и конца таблицы осуществляется из источника данных – упрощенным алгоритмом по сравнению с инструментом извлечения базы данных. Начало и конец таблицы, а также левую и правую границу также можно брать из альтернатив подполей других локаторов. По этим границам отсеиваются линий, перед построением сеток. В случае поиска многостраничной таблицы можно указать чтобы начало таблицы искалось только на первой странице. Начало таблицы не может найтись после конца.

Принудительно рисовать гор линии, достраивать вертикальные линии. В случае, когда шапка таблицы разлинована вертикально, а сама таблица – нет, можно использовать **достроение вертикальных линий**. Так же есть принудительное **достроение**

горизонтальных линий и настройки последующего объединения строк, это нужно, когда таблица не разлинована горизонтально. В случае поиска многостраничной таблицы может еще пригодится поиск промежуточного начала. Т.е. шапки таблицы на второй и последующей странице таблицы. Выполняется поиск начала таблице похожего на то что нашлось на первой странице. Это служит для отсеивания линий, которые могут быть между табличными данными и мешать правильному объединению сеток с разных страниц.



(Рис. 68 Локатор «Разлинованные таблицы» - Настройки без линий/линий Хафа)

Морфологические преобразования для поиска линий. Выполняется адаптивная бинаризация изображения в оттенках серого. Затем результат «разъедается» и «расширяется» с использованием структурного элемента – прямоугольник. В результирующе изображении проводится поиск контуров, и построение линий из подходящих. Настройки для метода не вынесены.

Минимальная ширина линии. Указывает минимальную ширину линии в обратных долях от ширины изображения.

Минимальная высота линии. Указывает минимальную высоту линии в обратных долях от высоты изображения.

Получение линий из OCR. В этом случае в качестве линий берутся текстовые линии, соответствующих типов из результатов ocr для поиска линий.

Получение линий исходя из структуры ocr. Служит для того чтобы создавать линии между колонками, строками таблицы, если таковые отсутствуют или не извлекаются другими методами. Анализируются размеры и координаты результатов ocr для определения отступов по вертикали и горизонтали, в которых создаются соответствующие линии. Имеющиеся настройки: максимальный вертикальный разрыв, доля максимальной строки, расстояние до горизонтальной линии, расстояние до вертикальной линии, доля длинной линии, отступ в таблице, ширина символа по умолчанию, количество пробельных символов.

Максимальный вертикальный разрыв. Максимальное расстояние по вертикали между соседними текстовыми линиями, при котором они могут войти в одну таблицу, выраженное в процентах от высоты изображения. Диапазон значений: 0-100.

Доля максимальной строки. Доля от максимальной длины текстовой строки, при которой она является первой строкой в табличной строке, если 0, то каждая текстовая строка будет отдельной строкой в таблице. Диапазон значений: 0-100.

Расстояние до горизонтальной линии. Максимальный процент разброса для координат по x слов при котором считается что слова начинаются/заканчиваются на одной вертикальной линии. Диапазон значений: 0-100.

Расстояние до вертикальной линии. Расстояние от координаты у строки в процентах от высоты страницы на котором надо создать линию. Диапазон значений: 0-100.

Доля длинной линии. Доля от количества "длинных" строк, при превышении которой определяется правая и левая граница таблицы. Диапазон значений: 0-100.

Отступ в таблице. Количество пикселей, на которое будут отличаться координаты таблицы в сторону ее расширения. Диапазон значений: 0-100.

Ширина символа по умолчанию. Ширина символа в таблице по умолчанию. Диапазон значений: 1-100.

Количество пробельных символов. Количество пробельных символов, которое говорит о том, что слова находятся в разных столбцах, при котором между словами необходимо провести вертикальную линию. Диапазон значений: 1-100.

Поиск линий по структуре изображения. Происходят преобразования изображения для выделения колонок таблиц, а затем в них строк. Между найденными столбцами и строками таблицы создаются линии. Настройки для метода не вынесены.

Минимальная длина линии. Указывается минимальная длина вертикальных линий.

Максимальная длина линии. Указывается минимальная длина горизонтальных линий.

Принудительно рисовать горизонтальные линии. Опционально создавать горизонтальные линии между всеми строками ocr который попали в область таблицы, не зависимо от того есть они там или нет.

Достраивать вертикальные линии. Опционально увеличивается длина вертикальных линий, если они пересекают (соприкасаются) верхнюю горизонтальную линию табличной сетки до нижней горизонтальной линии табличной сетки.

Разброс границ. Величина, показывающая на сколько процентов от ширины изображения, отодвинется найденная граница таблицы, влево (левая) или вправо (правая), если альтернатива, являющаяся маркером границы, находится не на рассматриваемой странице. Диапазон значений: 0-10.

Минимальное расстояние слева до таблицы. Указывает координату X, в процентах от ширины изображения, слева от которой не будет дорисовываться левая граница таблицы. Диапазон значений: 0-100.

Минимальная ширина столбца. Отсеивает столбцы, полученные из табличных сеток, если их ширина меньше указанного значения, выраженного в процентах от ширины изображения.

Отступ столбцов. Расстояние в пикселях, от линий в табличной сетке, от которого начинается построения столбца. Диапазон значений: 0-100.

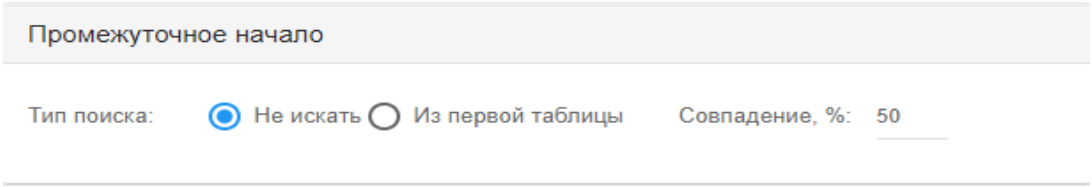
Отступ заполнения. Расстояние от левой и правой границы столбца, с которого начинается заполнение словами из ocr ячейки таблицы. Диапазон значений: -10 - 10.

Процент длины строки для объединения. При использовании объединения строк, количество столбцов в таблице умноженное на данное значение меньше чем количество

заполненных ячеек, тогда эта строка объединяется с предыдущей, если таковая имеется. Диапазон значений: 0-100.

Максимальный отступ от нижнего края для объединения. Расстояния в процентах от высоты листа от низа таблицы до донца листа, при превышении которого таблица считается законченной, т.е. к ней не прибавляется следующая. Диапазон значений: 0-100.

Максимальный отступ от верхнего края для объединения. Расстояния в процентах от высоты листа от верха страницы до начала таблицы на следующем листе, при превышении которого эта таблица не прибавляется к текущей, и текущая считается завершенной. Диапазон значений: 0-100.



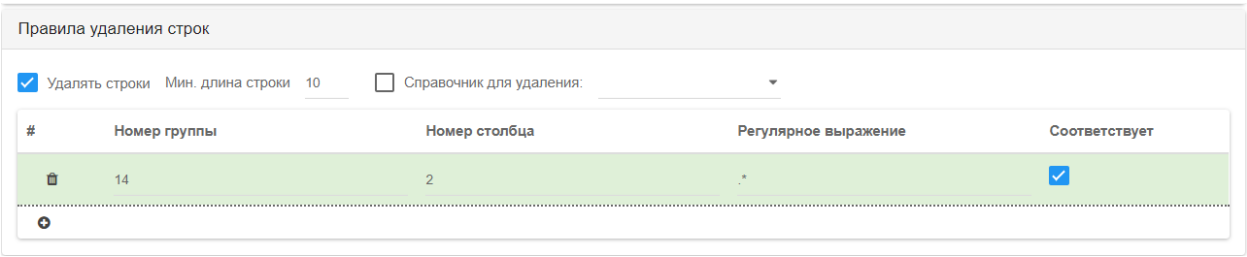
(Рис. 69 Локатор «Разлинованные таблицы» - Промежуточное начало)

Поиск промежуточного начала таблицы. Отвечает за поиск признаков начала таблицы на страницах, которые не являются первой страницей таблицы. Найденное промежуточное начало таблицы является локальным началом таблицы для каждой текущей страницы. Варианты: не искать, из первой таблицы, из источника.

Не искать. Если выбрана эта опция, то линий на страницах, последующих после первой страницы таблицы будут искаться от самого верха страниц.

Из первой таблицы. В этом случае, на страницах после первой страницы таблицы будет выполняться поиск начала таблицы, найденного на первой странице. Поиск будет осуществляться по неточному соответствию. Из полученных результатов будет выбрано лучшее по степени доверия.

Совпадение. Указывает на минимальный процент совпадения промежуточного начала таблицы с основным началом таблицы для его использования. Диапазон значений: 0-100.



(Рис. 70 Локатор «Разлинованные таблицы» - Правило удаления строк)

В таблицу часто могут попасть лишние данные, часть шапки, итоги и т.д. Для исключения этих данных есть правила удаления. Первое – это совпадения строки таблицы со строкой подключенного справочника более чем на 50%. Второе это длина текста в строке, если она меньше указанного значения, то она удаляется. Третье это Правила удаления, по совпадению или несовпадению содержимого указанного столбца с указанными регулярными выражениями. Например: в первом столбце должны быть только числа, или в третьем столбце не должно быть «Итого».

Удалять строки. Если эта опция включена, то к таблице после извлечения будут применены настройки удаления строк.

Минимальная длина строки. Указывает минимальное количество символов в строке таблицы при котором эта строка не будет удалена. Диапазон значений: 0-100.

Справочник для удаления. При использовании этой настройки, строки из указанного справочника будут сравниваться с текстом строк таблицы с помощью неточного соответствия, и в случае совпадения свыше 50% строка из таблицы будет удалена.

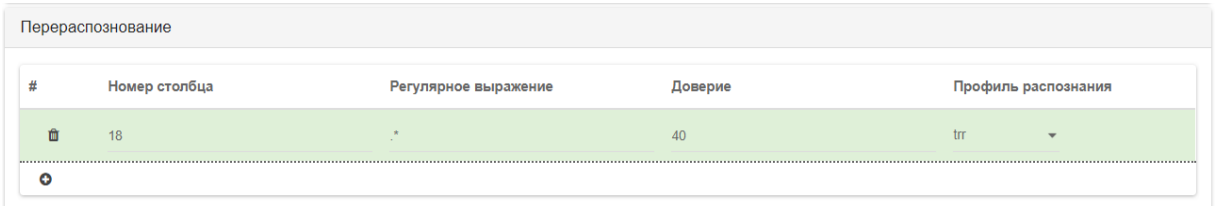
Правила удаления строк. Содержат коллекцию параметров для удаления строк из таблицы. Параметры: номер группы, номер столбца, регулярное выражение, соответствие.

Номер группы. Определяет группы правил, которые выполняются, при выполнении всех условий в группе.

Номер столбца. Определяет к какому столбцу будет применено указанное правило.

Регулярное выражение. Указывает регулярное выражение для проверки соответствия ему ячеек из указанного столбца.

Соответствует. При выбранной опции, строка будет удаляться если ячейка соответствует указанному регулярному выражению, при не выбранной опции – наоборот, если ячейка не соответствует регулярному выражению, произойдет удаление.



(Рис. 71 Локатор «Разлинованные таблицы» - Перераспознование)

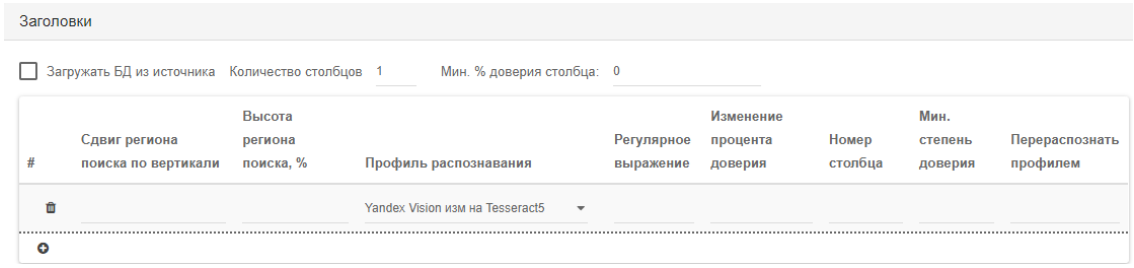
Настройки перераспознавания ячеек содержат номер столбца, минимальную степень доверия и регулярное выражение, при несоответствии с которым происходит перераспознавание ячейки указанным профилем распознавания. Можно использовать, например, для перечитывания ячеек с номером, количеством или ценой, когда известно, как должен выглядеть результат и можно использовать настройки «белого списка» символов при распознавании.

Номер столбца. Указывается для какого столбца применяются настройки.

Регулярное выражение. Указывает то регулярное выражение, при несоответствии текста ячейки которому, ячейка будет перераспознана.

Процент доверия. Указывает минимальный процент доверия к ячейке таблицы указанного столбца при котором не будет требоваться ее перераспознавание. Диапазон значений: 0-100.

Профиль распознавания. Указывает профиль распознавания, которым будет перераспознана ячейка, при выполнении соответствующих условий.



(Рис. 72 Локатор «Разлинованные таблицы» - Заголовки)

Настройки заголовков служат для тех таблиц, в которых необходимые столбцы могут располагаться в разной последовательности. При их использовании необходимо указать количество столбцов в итоговой таблице, минимальный процент доверия к столбцу, для помещения его в итоговую таблицу, а также указать перечень регулярных выражений, и зону относительно столбца, при попадании слова (при превышении им указанной степени доверия), соответствующего регулярному выражению в которую этот столбец на указанное количество процентов классифицируется как указанный. Что это значит? Например, в заголовке есть слова «Наименование» и «Цена». Указываем регулярное выражение «^Наименование$» для столбца №1 и «^Цена$» для столбца №2 и изменение процента доверия в 100 для обоих. Теперь независимо от того сколько каких столбцов есть в таблице, в итоговой будет 2 столбца, над которыми есть указанные слова. Если ни один из столбцов не подошел по условию, то в результате будет пустой столбец.

Загружать из БД. При включенной опции при выполнении инструмента извлечения таблицы с данными (для начала, конца таблицы и удаляемых строк) будут наполняться из файла или подключенной таблицы из базы данных. Иначе таблицы будут наполняться из временной копии, сохраненной в базе данных системы.

Количество столбцов. Количество столбцов в итоговой таблице при использовании заголовков.

Минимальный процент доверия столбца. Процент доверия, при превышении результатом заголовков которого, столбец будет записан в итоговую таблицу. Диапазон значений: 0-100.

Коллекция правил заголовков. Содержит настройки для формирования результатов поиска заголовков. Указанные параметры: сдвиг региона по вертикали, высота региона поиска, регулярное выражение, изменение процента доверия, номер столбца, минимальная степень доверия.

Сдвиг региона поиска. Указывает расстояние от верхней границы таблицы в процентах от высоты страницы на котором будет верхняя граница региона поиска признака столбца. Диапазон значений: -100 - 100. Т. е. при значении 0 значение будет искаться на уровне начала таблицы и ниже. Если указать отрицательное число (-10), то верзняя граница области поиска будет на 10% (10% от высоты листа) выше от начала страницы, если число положительное, то область поиска будет начинаться ниже начала таблицы.

Высота региона поиска. Указывает высоту региона поиска признака столбца, указанную в процентах от высоты страницы. Диапазон значений: 0 - 100. Т. е. высота области начиная от установленного сдвига относительно начала таблицы.

Левая граница региона поиска и его ширина определяется левой и правой границей столбца.

Регулярное выражение. Указывает регулярное выражение, которому должно соответствовать слово в указанном регионе, для того чтобы записать его в результаты поиска заголовков.

Изменение процента доверия. Указывает величину, на которую меняется итоговая степень доверия классификации столбца. Диапазон значений: -100 - 100.

Номер столбца. Указывает на результат по какому столбцу будет влиять правило заголовка.

Минимальная степень доверия. Указывает минимальную степень доверия к слову из ocr, чтобы оно могло попасть в результаты поиска заголовков.

Перераспознать профилем. Если профиль указан, то область заголовка будет перераспознана для повторного поиска в ней ключей заголовков, если они не были найдены в исходной реперезентации.

Подполя инструмента извлечения:

Таблица. Подполе содержащее таблицу, строки и столбцы.

Линии. Все линии внутри таблицы. Из найденных линий образуется сетка таблицы.

Колонки. Колонки найденной таблицы вместе с заголовками и шапкой.

Регионы заголовков. Области в которых происходит поиск заголовков.

Начало таблицы. Указывает альтернативу начала таблицы, найденную по справочнику или унаследованную с подполя.

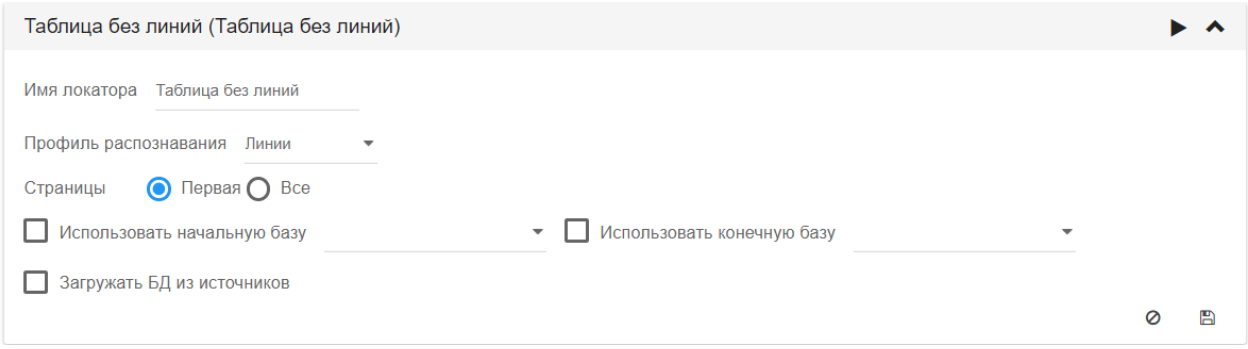
Конец таблицы. Указывает альтернативу конца таблицы, найденную по справочнику или унаследованную с подполя.

Левая граница таблицы. Указывает альтернативу левой границы таблицы, унаследованную с подполя.

Правая граница таблицы. Указывает альтернативу правой границы таблицы, унаследованную с подполя.

7. Таблица без линий ()

Частный случай табличного локатора, который позволяет искать линий лишь 4ым способом без настроек, имеет 1 профиль распознавания в настройках. Начало и конец таблицы ищет только по справочникам.



(Рис.73 Локатор «Таблица без линий»)

Профиль распознавания. Указывает репрезентацию, в которой будет осуществлен поиск таблицы.

Страницы, на которых будет выполнятся инструмент извлечения. Возможные варианты – первая страница. При поиске на **первой странице**, будет осуществляться поиск одностраничной таблицы только в указанной репрезентации первой страницы документа.

Источник данных начала таблицы. Указывает таблицу, значения строк которой будет сравниваться со строками в указанной репрезентации для поиска слов, указанных страниц. В качестве начала таблицы выбирается строка результатов ocr которая наиболее совпадает со строкой в ocr, в случае если процент совпадения строк выше 80. Может быть не указан.

Источник данных конца таблицы. Указывает таблицу, значения строк которой будет сравниваться со строками в указанной репрезентации для поиска слов, указанных страниц. В качестве конца таблицы выбирается строка результатов ocr которая наиболее совпадает со строкой в ocr, в случае если процент совпадения строк выше 80. Поиск осуществляется среди строк, находящихся после (ниже в ocr и на последующих страницах), начала таблицы, если таковое выло найдено. Может быть не указан.

Загружать данные из источника. При включенной опции при выполнении инструмента извлечения таблицы с данными (для начала и конца таблицы) будут наполняться из файла или подключенной таблицы из базы данных. Иначе таблицы будут наполняться из временной копии, сохраненной в базе данных системы.

Подполя инструмента извлечения:

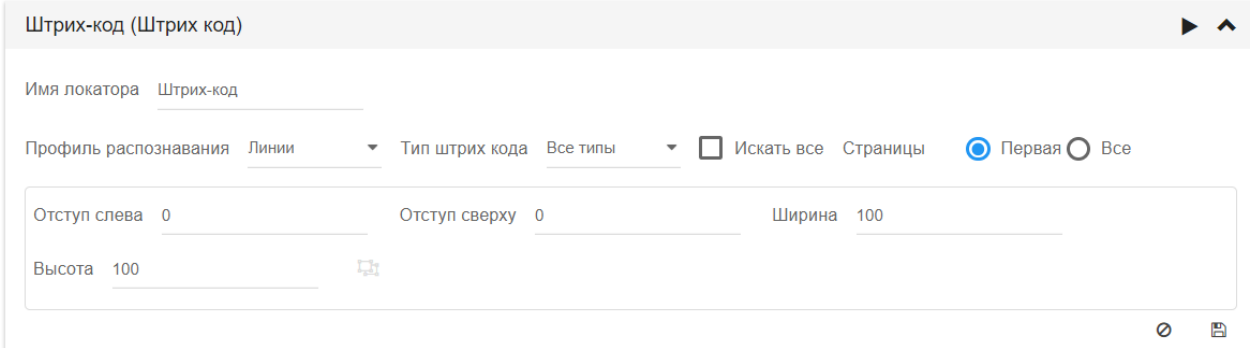
Таблица. Подполе содержащее таблицу, строки и столбцы.

Начало таблицы. Указывает альтернативу начала таблицы, найденную по справочнику.

Конец таблицы. Указывает альтернативу конца таблицы, найденную по справочнику.

8. Штрих-код ( )

Служит для поиска штрих кода. Поиск осуществляется на указанной репрезентации первой или всех страниц документа, в указанном регионе. Можно указать конкретный тип штрих кода для поиска, или искать любого типа. Так же есть опция поиск всех штрих кодов, при которой создается несколько альтернатив единственного выходного подполя.



(Рис. 74 Локатор «Штрих-код»)

Профиль распознавания. Указывает репрезентацию, в которой будет осуществлен поиск штрих кодов.

Настройки области инструмента извлечения. Описывает область репрезентации, в которой будет выполняться поиск штрих кодов. Можно указать числовое значение, либо графически на репрезентации.

- **Отступ слева** – отступ от левой границы изображения репрезентации выраженный в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.
- **Отступ сверху** – отступ от верхней границы изображения репрезентации выраженный в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.
- **Ширина** – ширина области, выраженная в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.
- **Высота** – высота области, выраженная в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.

Суммы Отступ слева, Ширина и Отступ сверху, Высота так же не могут быть больше 100.

Страницы, на которых будет выполнятся инструмент извлечения. Возможные варианты – первая страница, все страницы. При поиске на **первой странице**, будет осуществляться поиск штрих кодов на изображениях репрезентаций указанного профиля только первой страницы документа. При поиске на **всех страницах**, будет осуществляться поиск штрих кодов на изображениях репрезентаций указанного профиля на каждой из страниц документа.

Тип штрих кода. Указывает формат штрих кода, которые необходимо найти на изображении страницы. Варианты форматов: "Все типы", "Code128", "Code39", "EAN13", "EAN8", "PDF417", "QRCode", "UPCA", "UPCE", "ITF", "DataMatrix". https://ru.wikipedia.org/wiki/Сравнение_характеристик_штрихкодов

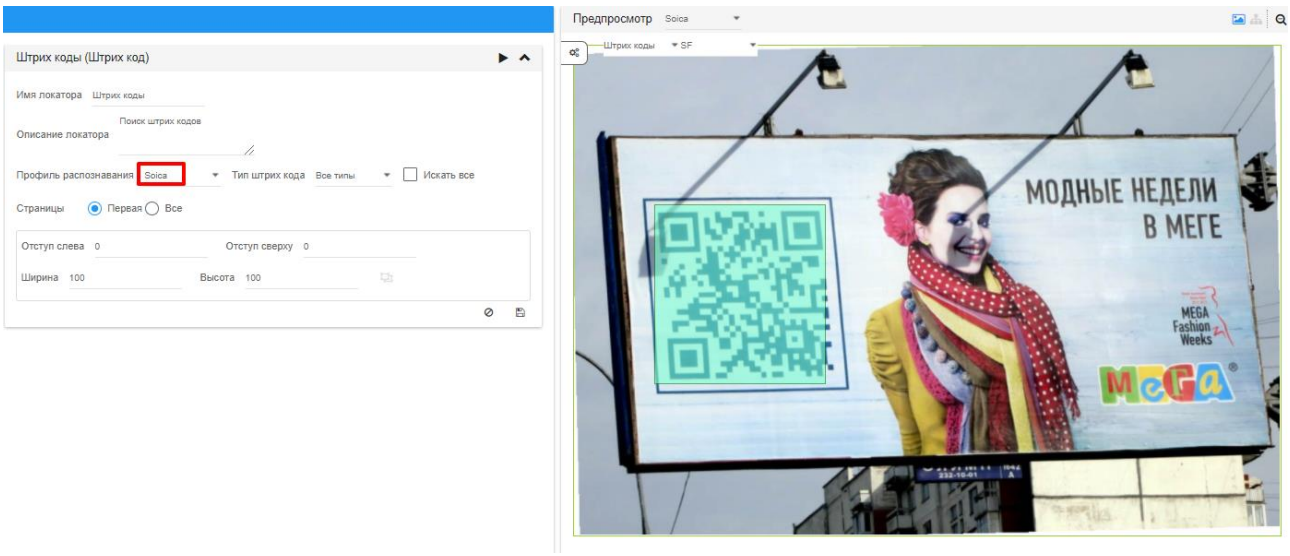
Искать все штрихкоды. При указании этой опции поиск штрих кода не будет ограничиваться первым попавшимся вариантом.

Подполя инструмента извлечения:

Результат. Подполе, содержащее альтернативы с найденными штрих кодами, где в качестве текста – информация из штрих кода.

Элементная база движка SOICAII также может использоваться в локаторе Штрих ходы.

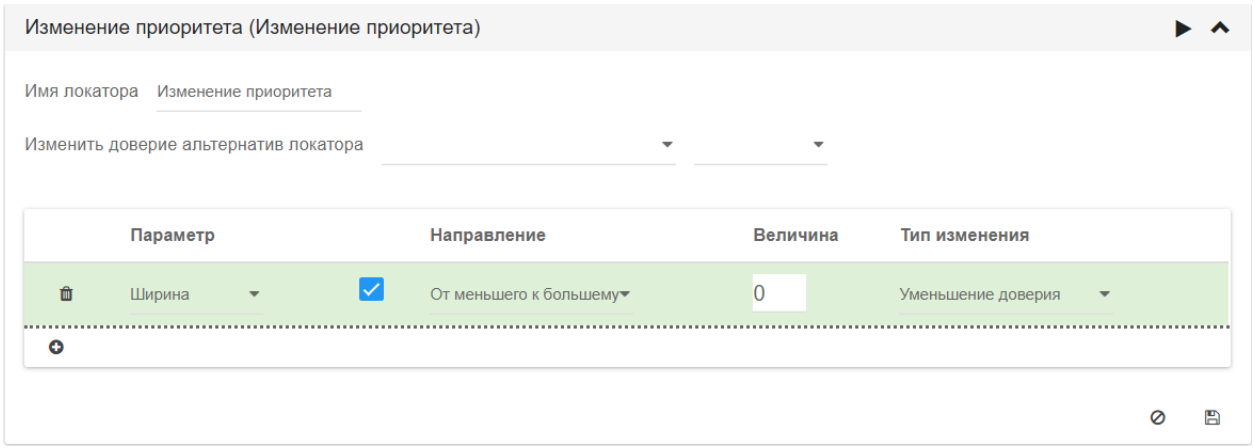
Для использования элементов движка SOICAII в этом локаторе необходимо выбрать профиль распознавания, в котором используется движок SOICAII с выбранной опцией поиска штрих кодов и убедиться, что в выбранном профиле распознавания на базе движка SOICAII выбрана соответствующая опция.



(Рис. 74.1 Настройка и результат локатора Штрих коды)

9. Изменение приоритета ( )

Локатор позволяет изменять степень доверия наследуемых альтернатив. Признаки, которые можно использовать для изменения степени доверия: номер страницы, длина текста, координата x, координата y, ширина, высота. Степень доверия можно увеличивать, уменьшать, или распределить (при этом текущая степень доверия альтернативы не учитывается). Так же задается направление изменения (от большего к меньшему или, от меньшего к большему). Кроме этого задается процент изменения доверия по каждому из параметров. Можно использовать, например, когда необходимы результаты с последней страницы, в этом случае нужно выбрать параметр номер страницы, выбрать уменьшение доверия от большего к меньшему и установить процент на 100. Так же можно пересортировать альтернативы по координатам, используя параметры X и Y, с процентом – 50, выравниванием степени доверия, от меньшего к большему.



(Рис. 75 Локатор «Изменение приоритета»)

Изменить доверие альтернатив локатора. Указывает подполе локатора, содержащее альтернативы, степень доверия которых будет изменяться.

Параметры изменения доверия альтернатив. Список настроек, которые позволяют влиять на степень доверия альтернатив наследуемого подполя локатора. Среди них: тип

параметра, использовать параметр, направление изменения, величина изменения, тип изменения.

Тип параметра. Указывает на тип свойства, в зависимости от значения которого будет изменяться степень доверия. Варианты: номер страницы, ширина, высота, X, Y, длина текста.

- **Номер страницы**. Номер страницы в документе, на которой расположена альтернатива.
- **Ширина**. Ширина региона альтернативы.
- **Высота**. Высота региона альтернативы.
- **X**. Отступ от левого края изображения региона альтернативы.
- **Y**. Отступ от верхнего края изображения региона альтернативы.
- **Длина текста**. Количество символов в тексте альтернативы.

Использовать параметр. Если эта опция включена, то значение указанного свойства будет влиять на степень доверия альтернатив наследуемого локатора.

Направление изменения. Указывает направление, в котором будет рассчитываться градиент изменение степени доверия альтернатив. Варианты: от меньшего к большему и от большего к меньшему.

Величина изменения. Указывается максимальное значение, на которое изменяется доверие альтернативы по указанному параметру. Диапазон значений: 0-100.

Тип изменения. Указывает в какую сторону происходит изменение доверия альтернативы по указанному параметру. Варианты: уменьшение, увеличение, уменьшение и увеличение.

- **Уменьшение доверия**. В этом случае степень доверия альтернатив, будет уменьшаться от текущего значения.
- **Увеличение доверия**. В этом случае степень доверия альтернатив, будет увеличиваться от текущего значения.
- **Уменьшение и увеличение доверия**. В этом случае степень доверия альтернатив сначала усредняется, а затем, будет равномерно распределяться в рамках указанной границы (величина изменения).

Подполя инструмента извлечения:

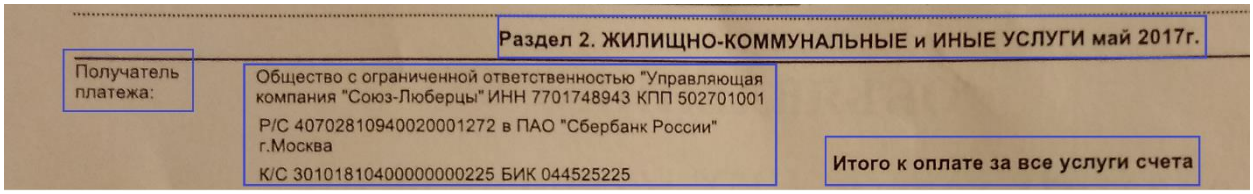
Результат. Подполе, содержащее альтернативы наследуемого подполя локатора с измененной степенью доверия.

10. Объединение ()

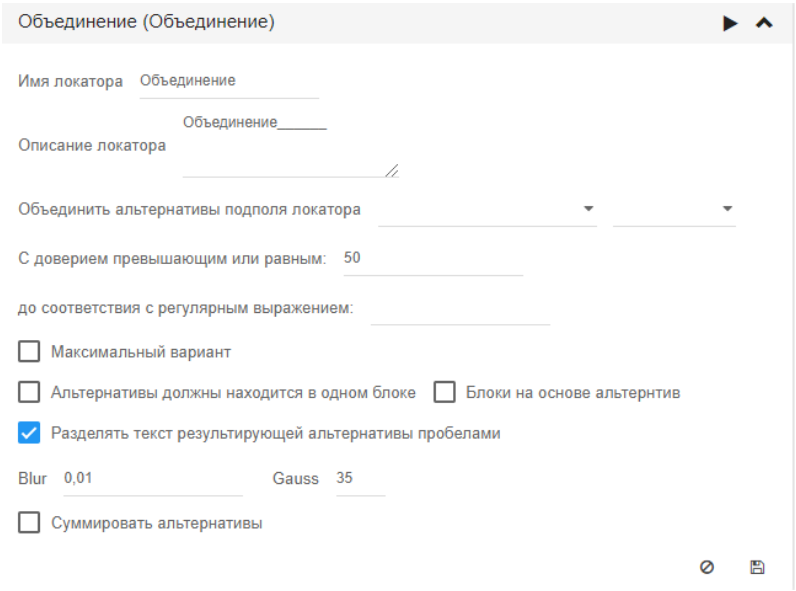
Позволяет объединять альтернативы одного подполя локатора. Объединяемые альтернативы можно ограничить минимальной степенью доверия, а также объединение можно завершать при достижении объединенной альтернативой совпадения с указанным регулярным выражением. Т.е., например, на входе альтернативы: «1», «2», «3», «п», «4», «5», «р», «6» и регулярное выражение «^\d\s\d$», в этом случае в результате будут следующие альтернативы: «1 2», «2 3», «4 5». И это в том случае, если будет выбрана опция ставить пробелы между альтернативами.

Так же можно использовать эти правила объединения только в том случае, если альтернативы входят в один абзац. Абзац вычисляется на основе всего изображения или

указанных альтернатив графическим способом по скоплениям объектов. Имеются настройки для корректировки определения абзацев. Это позволяет, выполнять, объединение альтернатив по регулярному выражению идущих не подряд с точки зрения порядка: строки/слова. Т.е. при поиске абзацев с помощью регулярного выражения: ^(000|ООО|ЗАО|ОАО|АО|ПАО|Общество\sс\sограниченной\sответственностью|(Закрытое|Открытое|Публичное)?\s?[а|А]кционерное\sобщество)\s([^(\"\b)]*\s)?["|“|'|”|‘][^(\"\b)]*["|“|'|”|‘].*?$ можно найти «Общество с ограниченной ответственностью "Управляющая компания "Союз-Люберцы"» и ПАО «Сбербанк России», если в качестве входного подполя будет AdvancedFormat инструмент извлечения локатор расширенного регулярного выражения с регулярным выражением «.*».



(Рис. 76 Пример работы локатора «Объединение»)



(Рис. 77 Локатор «Объединение»)

Наследуемое подполя инструмента извлечения. Указывает подполе локатора, содержащее альтернативы, которые будут объединяться данным инструментом извлечения.

Минимальная степень доверия. Указывает минимальную степень доверия альтернатив наследуемого локатора при которой альтернатива может участвовать в объединении. Диапазон значений: 0-100.

Регулярное выражение. Опциальное регулярное выражение, при достижении соответствия которому из объединяемых альтернатив создается альтернатива текущего локатора.

Альтернативы должны находится в одном блоке. Опция, позволяющая ограничить объединение альтернатив пределами графически найденной области скопления текста (абзаца или текстового блока).

Блоки на основе альтернатив. В этом случае, текстовые блоки формируются исходя только из регионов альтернатив наследуемого подполя локатора, а не исходя из графической структуры всего изображения, как с отключенной опцией.

Разделять текст результирующей альтернативы пробелами. При выборе этой опции все альтернативы наследуемого локатора будут занесены в результирующую альтернативу через пробел. После последней альтернативы так же будет добавлен пробел.

Blur (Ядро размытия). Указывает размер блока для адаптивной бинаризации и обычного размытия, выраженный в процентах от размера диагонали изображения.

Gauss (Ядро Гаусса). Указывает отклонение по X для фильтра Гаусса, в пикселях.

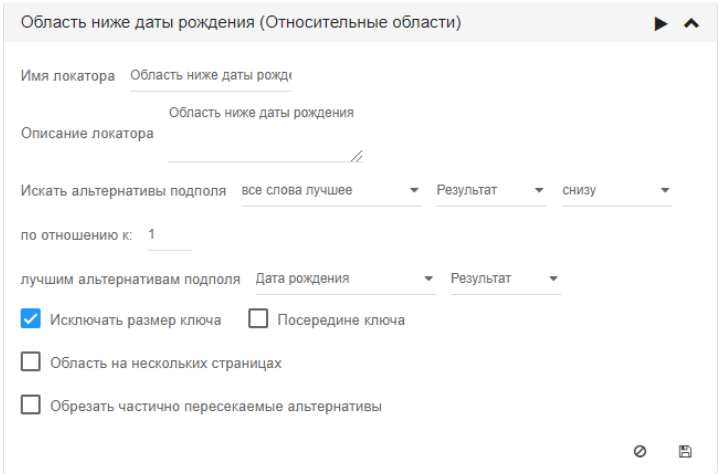
Суммировать альтернативы. При выборе этого параметра будет производиться автосумма,

Подполя инструмента извлечения:

Результат. Подполе, содержащее объединенные альтернативы наследуемого подполя.

11. Относительные области ( )

По сути это тоже самое что локатор «область ключевого слова, но в данном случае область указывается не в долях от ключа, а слева, справа, сверху, снизу ключа. С остальных сторон область ограничивается границами изображения. Кроме того, есть опция включать ли в область сам ключ.



(Рис. 78 Локатор «Относительные области»)

Имя локатора. Редактируется имя выбранного локатора.

Искать альтернативы подполя. Подполе локатора, альтернативы которого записываются в результирующее подполе, при попадании их в указанные области.

Лучшим альтернативам подполя. Подполе локатора, указанное количество лучших (по степени доверия) альтернатив которого образуют ключи, относительно которых рассчитываются ключевые области поиска.

По отношению. Указывает количество лучших (по степени доверия) альтернатив ключевого инструмента извлечения, которые образуют ключи.

Направление формирование области ключа. Указывает с какой стороны от области ключа будет проходить одна из границ ключевой области. Варианты: слева, справа, сверху, снизу.

Слева. В этом случае область ограничивается левой, верхней, нижней границей изображения, а справа – областью ключа.

Справа. В этом случае область ограничивается правой, верхней, нижней границей изображения, а слева – областью ключа.

Сверху. В этом случае область ограничивается правой, левой, верхней границей изображения, а снизу – областью ключа.

Снизу. В этом случае область ограничивается правой, левой, нижней границей изображения, а сверху – областью ключа.

Исключать размер ключа. При выбранной опции граница области, ограниченная ключом, будет проходить по той границе ключа, что ближе к противоположной границе области, если опция не выбрана – то область самого ключа будет включена в конечную область поиска.

Подполя инструмента извлечения:

Результат. Альтернативы наследуемого подполя инструмента извлечения, которые полностью попали в одну или несколько областей рассчитанных, из указанного количества альтернатив ключевого подполя инструмента извлечения по их координатам и настройкам расчета области применения ключа.

Ключи – указанное количество лучших (по степени доверия) альтернатив ключевого подполя инструмента извлечения.

Области – области ключа для найденных ключей, рассчитанных исходя из размера и расположения ключа и настроек расчета области применения ключа.

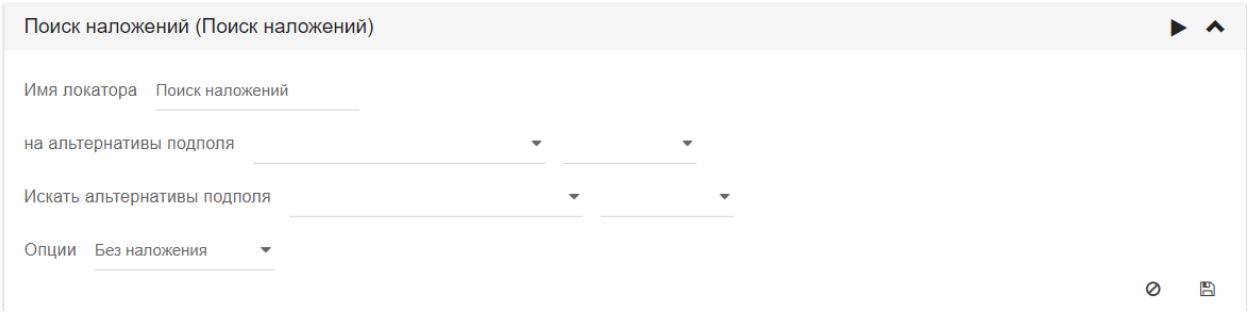
Область на нескольких страницах. Выделяется область, начинающаяся на одной странице и заканчивающаяся на другой. Пример: таблица в ТТН начинается на одной странице, заканчивается на другой. При выборе данного параметра в область попадет вся таблица в ТТН.

Обрезать частично пересекаемые альтернативы. Если при выборе данного параметра, указано, например, «ниже», а альтернатива находится и ниже и выше ключа, то в результат попадет только часть альтернативы, которая находится ниже.

12. Поиск наложений ()

Служит для поиска альтернатив одного подполя, которые либо не накладываются ни на одну альтернативу второго подполя, либо частично накладываются, либо полностью попадают в одно из них. В зависимости от выполнения условия создаются альтернативы с текстом True или False, содержащие остальные параметры из наследуемых альтернатив.

Условно говоря локатор проверяет попадает ли одна альтернатива в область другой альтернативы и передает значение True или False.



(Рис. 79 Локатор «Поиск наложений»)

Имя локатора. Редактируется имя выбранного локатора.

на альтернативы подполя. Подполе, альтернативы которого проверяются на геометрическое соответствие с альтернативами подполя подложки.

Искать альтернативы подполя. Подполе, содержащее альтернативы, с которыми сравниваются геометрические параметры подполя накладывающихся альтернатив.

Опции. Тип условия, при котором накладывающиеся альтернативы, попадут в результирующее подполе инструмента извлечения с текстом True. Варианты: без наложения, полное наложение, частичное наложение.

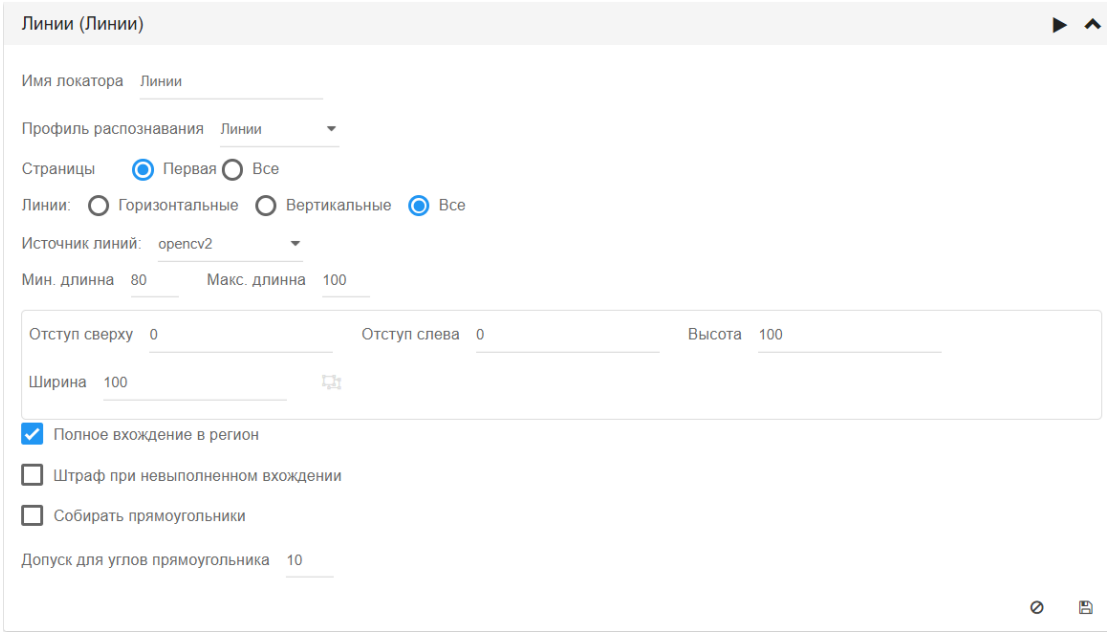
- **Без наложения**. Накладывающаяся альтернатива попадет в результирующе подполе с текстом True, в том случае, если она не пересекает ни одну альтернативу из поля подложки.
- **Полное наложение.** Накладывающаяся альтернатива попадет в результирующе подполе с текстом True, в том случае, если она попадает полностью хотя бы в одну альтернативу из поля подложки.
- **Частичное наложение**. Накладывающаяся альтернатива попадет в результирующе подполе с текстом True, в том случае, если она попадает хотя бы частично хотя бы в одну альтернативу из поля подложки.

Подполя инструмента извлечения:

Результат. Альтернативы, из подполя накладывающихся альтернатив, с текстом True или False, в зависимости от того выполнены ли заданные условия.

13. Линии ( )

Выполняет поиск горизонтальных и вертикальных линий одним из 3х способов поиска линий/ Поиск выполняется в указанном регионе указанной репрезентации первой или всех страниц документа. В настройках есть минимальная и максимальная длина линии относительно высоты или ширины изображения. Также есть опция, при которой в результаты записываются только те линии, которые полностью попали в указанный регион. Так же можно выбрать опцию, при которой степень доверия к альтернативе будет равна проценту попадании линии в регион. Найденные линии имеют толщину 6 пикселей и записываются в альтернативы единственного подполя с текстом “v” (вертикальные) или “h” (горизонтальные).



(Рис. 80 Локатор «Линии»)

Профиль распознавания. Указывает репрезентацию, в которой будет осуществлен поиск линий.

Настройки области инструмента извлечения. Описывает область репрезентации, в которой будет выполняться поиск линий. Можно задать графически на репрезентации.

- **Отступ слева –** отступ от левой границы изображения репрезентации выраженный в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.
- **Отступ сверху** – отступ от верхней границы изображения репрезентации выраженный в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.
- **Ширина –** ширина области, выраженная в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.
- **Высота –** высота области, выраженная в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.

Суммы Отступ слева, Ширина и Отступ сверху, Высота так же не могут быть больше 100.

Страницы, на которых будет выполнятся инструмент извлечения. Возможные варианты – первая страница, все страницы. При поиске на **первой странице**, будет осуществляться поиск линий на репрезентациях указанного профиля только первой страницы документа. При поиске на **всех страницах**, будет осуществляться поиск линий на репрезентациях указанного профиля на каждой из страниц документа.

Тип линии. Указывает какие типы линий будут искаться инструментом извлечения. Варианты: горизонтальные, вертикальные, все.

Источник линий. Указывает алгоритм, которым будет производится поиск линий. Варианты: метод Хафа, Морфологические преобразования, OCR.

Метод Хафа для поиска линий. В этом случае происходит преобразование Кэнни размытого изображения, переведенного в оттенки серого, в результатах которых выполняется поиск линий методом Хафа. Настройки для метода не вынесены.

Морфологические преобразования для поиска линий. Выполняется адаптивная бинаризация изображения в оттенках серого. Затем результат «разъедается» и «расширяется» с использованием структурного элемента – прямоугольник. В результирующе изображении проводится поиск контуров, и построение линий из подходящих. Настройки для метода не вынесены.

Получение линий из OCR. В этом случае в качестве линий берутся текстовые линии, соответствующих типов из результатов ocr для поиска линий.

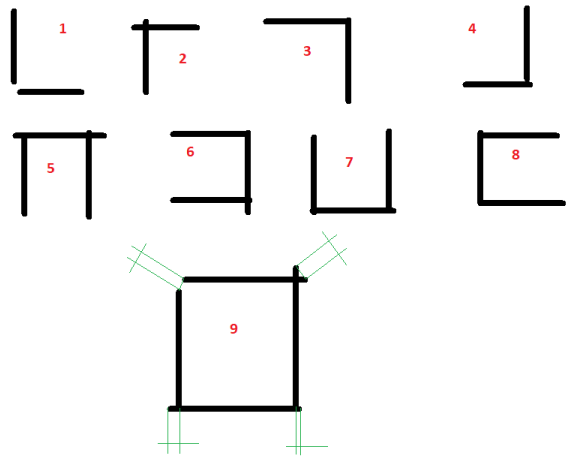
Минимальная длина линии. Выполняется отсеивание линий, которые меньше по длине чем указанное значение. Указывается в процентах от высоты (вертикальные линии) или ширины (горизонтальные линии) изображения. Диапазон значений: 0-100.

Максимальная длина линии. Выполняется отсеивание линий, которые меньше по длине больше чем указанное значение. Указывается в процентах от высоты (вертикальные линии) или ширины (горизонтальные линии) изображения. Диапазон значений: 0-100.

Полное вхождение в регион. При выборе этой опции, линии, область которых не полностью попадает в область поиска, не будут записываться в локатора.

Штраф при неполном вхождении. При выборе этой опции, степень доверия линии будет рассчитываться исходя из того, какой процент ее области попал в регион поиска линий.

Собирать прямоугольник. Если выбрана эта опция, то найденные линии будут пробовать собираться в фигуры следующих типов:



Допуск для углов прямоугольника (9)– это максимальное расстояние между концами линий, при которым считается что они образуют угол.

Подполя инструмента извлечения:

Результат. Альтернативы содержащие горизонтальные и/или вертикальные линии, с текстом «h» и «v», соответственно.

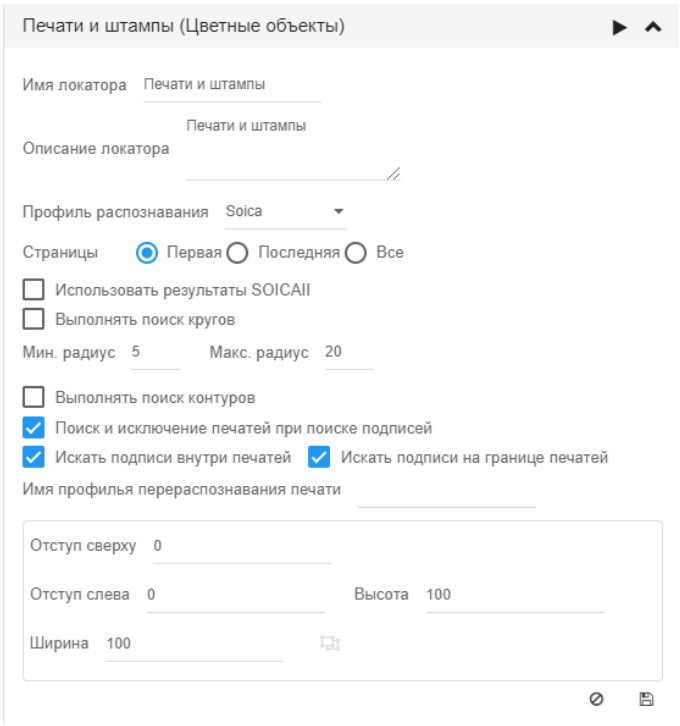
14. Печати и штампы ( )

Локатор выполняет поиск линий, кругов и рукописного текста на изображении. Локатор выполняет поиск кругов (печатей) и на бинаризированом изображении. Печать или рукописный текст не обязательно должны быть цветными.

Предварительно изображение может фильтроваться по каждой из компонентов цветовой модели hsv. Эти компоненты – Оттенок, Насыщенность и Яркость. Можно задать диапазоны на каждый компонент и использовать сочетание отфильтрованных компонентов. На выходе каждого канала – монохромное изображение. Например, для фильтрации синих и фиолетовых чернил подойдут настройки: h: 180-300, s: 15-100, v: 20-80.

Альтернативами локатора являются все найденные в указанном регионе поиска круги и области изображения с рукописным текстом.

Внутри печати можно искать текст.



(Рис. 81 Локатор «Цветные объекты» - Базовые настройки)

Профиль распознавания. Указывает репрезентацию, в которой будет осуществлен поиск объектов.

Настройки области инструмента извлечения. Описывает область репрезентации, в которой будет выполняться поиск объектов.

- **Отступ слева** – отступ от левой границы изображения репрезентации выраженный в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.
- **Отступ сверху** – отступ от верхней границы изображения репрезентации выраженный в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.

- **Ширина** – ширина области, выраженная в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.
- **Высота** – высота области, выраженная в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.

Суммы Отступ слева, Ширина и Отступ сверху, Высота так же не могут быть больше 100.

Страницы, на которых будет выполнятся инструмент извлечения. Возможные варианты – первая страница, все страницы. При поиске на **первой странице**, будет осуществляться поиск объектов на репрезентациях указанного профиля только первой страницы документа. При поиске на **всех страницах**, будет осуществляться поиск объектов на репрезентациях указанного профиля на каждой из страниц документа.

Поиск кругов. При включенной опции будет производится поиск кругов в отфильтрованном изображении. Включает настройки: минимальный радиус, максимальный радиус.

Минимальный радиус. Радиус, меньше которого круги не будут найдены, выраженный в процентах от максимального размера изображения. Диапазон значений: 1-100.

Максимальный радиус. Радиус, больше которого круги не будут найдены, выраженный в процентах от максимального размера изображения. Диапазон значений: 1-100.

Выполнять поиск контуров. При включенной опции осуществляет поиск рукописного текста на изображении.

Возможна настройка совместного поиска подписей с печатями: исключение печатей,

поиск внутри печатей, поиск на границе печатей. Для этого необходимо отметить нужные опции галочкой.

Имя профиля перераспознавания печати. Печать разворачивается и пытается прочитаться выбранным для перераспознавания профилем.

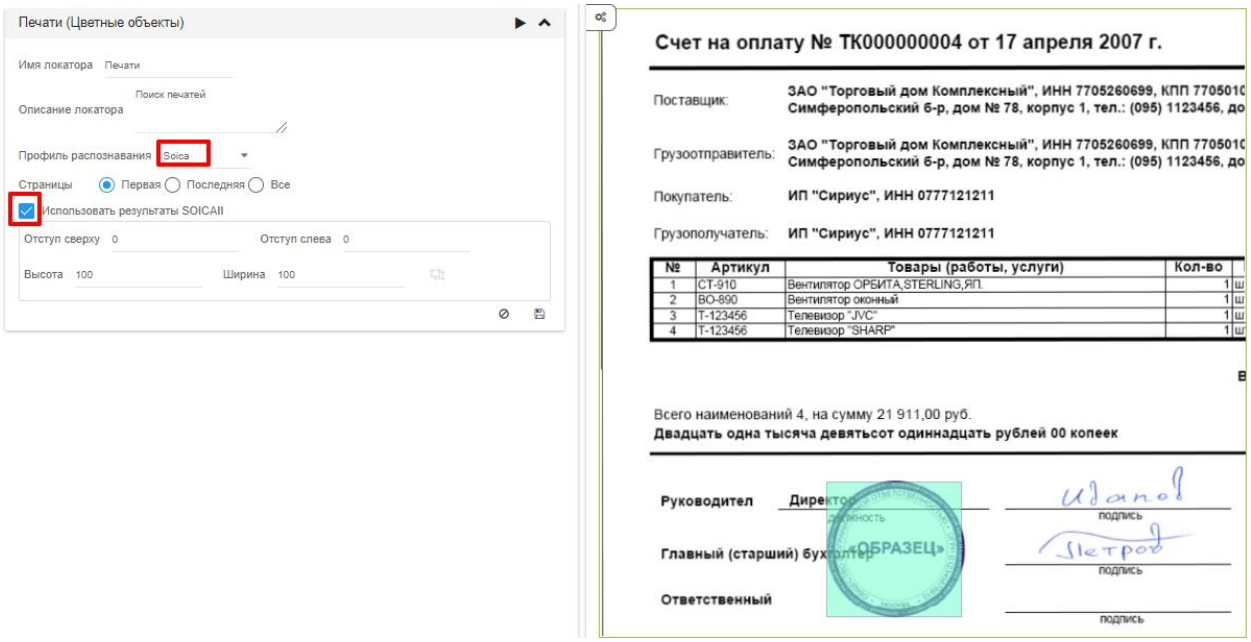
Подполя инструмента извлечения:

Круги. Альтернативы содержащие найденные круги с текстом True.

Линии. Альтернативы содержащие горизонтальные и/или вертикальные линии, с текстом «h» и «v», соответственно.

Подписи. Альтернативы содержащие рукописный текст, с текстом sign.

Для использования элементов движка SOICAII в этом локаторе необходимо выбрать профиль распознавания с движком SOICAII, отметить галочкой опцию «Использовать результаты SOICAII» и убедиться, что в выбранном профиле распознавания на базе движка SOICAII выбрана соответствующая опция.

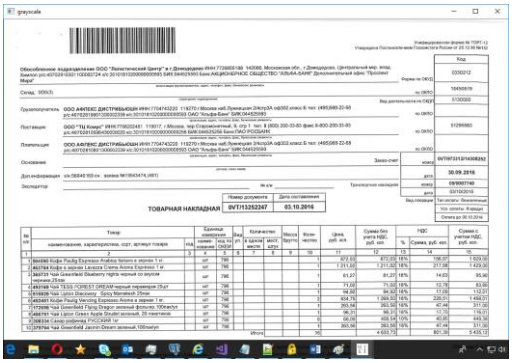


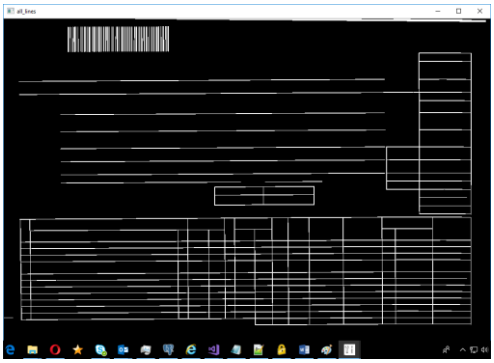
(Рис. 81.1 Настройка и результат локатора Печати и штампы)

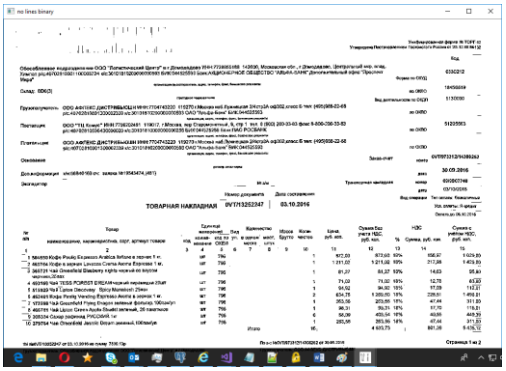
15. Абзац ( )

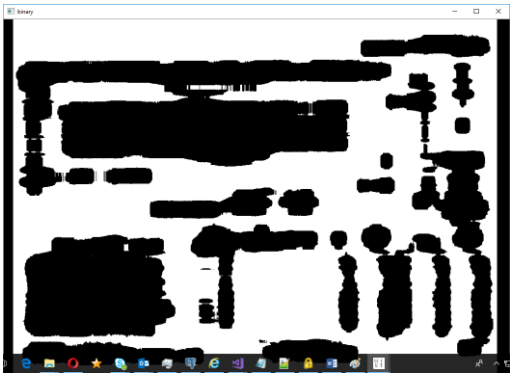
Локатор группирует графические объекты, найденные на изображении и делает из полученных групп альтернативы или фильтровать слова по определенному признаку или группе признаков.

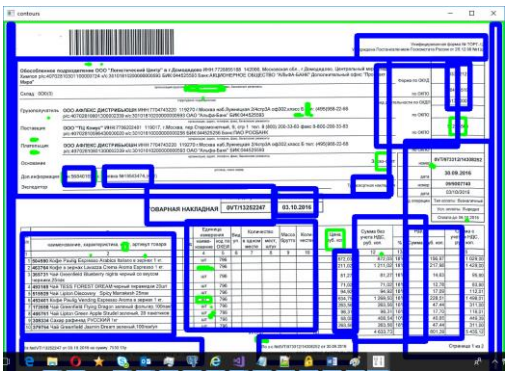
Пример работы локатора:

1. 

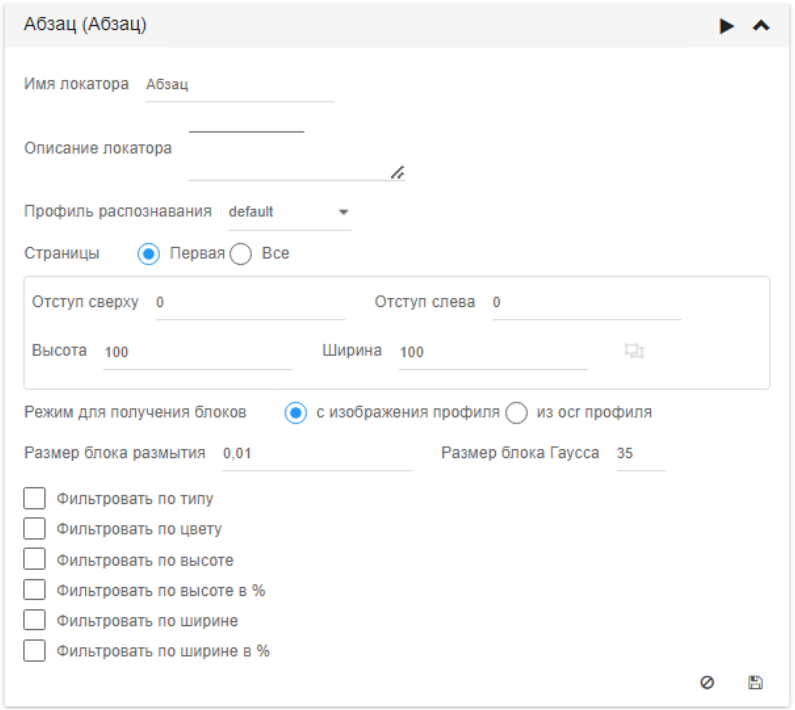
2. 

3. 

4. 

5. 

1. Оригинал; 2. Нахождение линий; 3. Удаление линий; 4. Объединение объектов; 5. Нахождение и отсеивание контуров полученных объектов.



(Рис. 86 Локатор «Абзац»)

Профиль распознавания. Указывает репрезентацию, в которой будет осуществлен поиск объектов.

Страницы, на которых будет выполнятся инструмент извлечения. Возможные варианты – первая страница, все страницы. При поиске на **первой странице**, будет осуществляться поиск объектов на репрезентациях указанного профиля только первой страницы документа. При поиске на **всех страницах**, будет осуществляться поиск объектов на репрезентациях указанного профиля на каждой из страниц документа.

Настройки области инструмента извлечения. Описывает область репрезентации, в которой будет выполняться поиск объектов.

- **Отступ слева** – отступ от левой границы изображения репрезентации выраженный в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.

- **Отступ сверху** – отступ от верхней границы изображения репрезентации выраженный в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.
- **Ширина** – ширина области, выраженная в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.
- **Высота** – высота области, выраженная в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.

Суммы Отступ слева, Ширина и Отступ сверху, Высота так же не могут быть больше 100.

Режим для получения блоков. Выбор режима и принципа объединения объектов на изображении. В режиме использования OCR в блоки объединяются только слова из репрезентации, иначе – любые графические объекты.

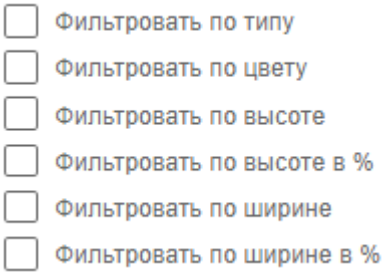
Размер блока размытия. – измеряется в долях от диагонали изображения (0,03 это 3%).

Размер блока Гаусса – измеряется в точном значении.

Подполя инструмента извлечения:

Абзацы – сгруппированные по заданным правилам области на изображении, с содержащимся в них ocr.

Ниже представлена группа настроек чтобы фильтровать слова по определенному признаку или группе признаков.

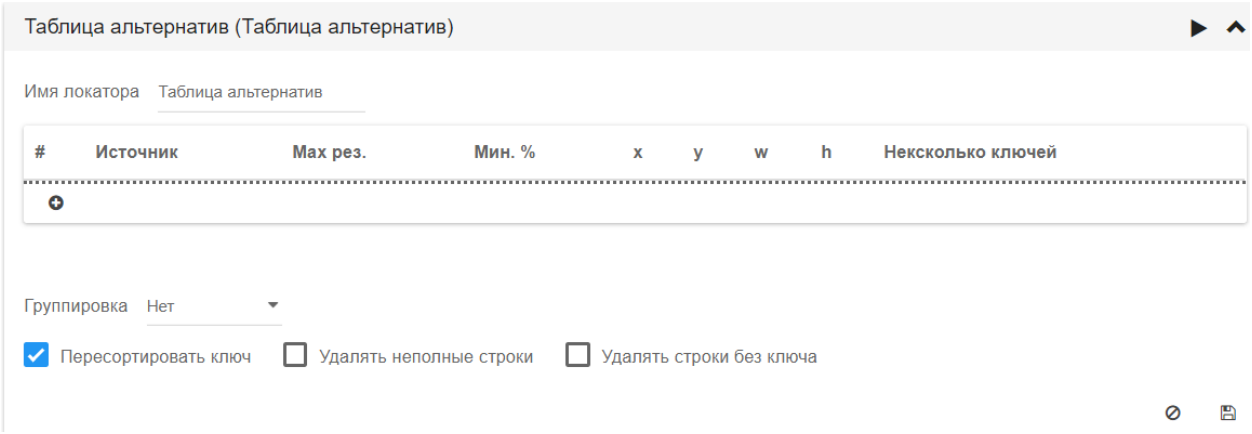


16. Таблица альтернатив ( )

Служит для объединения альтернатив подполей локаторов в таблицы. Без группировки альтернативы из каждого указанного подполя, ограниченные максимальным количеством и минимальной степенью доверия записываются в столбцы результирующей таблицы. При простой группировке, значения из столбцов пересортировываются по принципу ближайшего геометрического расположения альтернатив со второго и последующего столбцов к первому ключевому. Расширенная группировка позволяет группировать альтернативы по строкам таблицы по принципу попадания альтернатив 2 и последующего столбца в относительные области от первого столбца (ключа). При этом опционально можно: позволить одной альтернативе попадать в несколько ключевых областей или только в одну; позволить нескольким альтернативам неключевого столбца сгруппироваться с одним ключевым столбцом (при этом создастся несколько строк) или объединять альтернативы из неключевого столбца попавшие в зону ключа. В получившейся таблице можно: выполнять сортировку записей по степени доверия ключевого поля

(первого столбца); удалять строки, не содержащие ключевого столбца или имеющие хотя бы один пустой столбец.

Пример использования. Есть анкета в который справа от поле для галочки (чекбокса) указан артикул. Необходимо получить список выбранных артикулов. Выполняем поиск всех чекбоксов, также выполняем поиск артикулов. С помощью локатора форматирования изменяем степень доверия неотмеченных чекбоксов. В локаторе устанавливаем минимальную степень доверия для ключа, которым выбрать отформатированные чекбоксы, в качестве второго столбца выбираем артикулы. Выбираем расширенную группировку. Выбираем опцию попадания только лучшей альтернативы неключевого столбца. Удаляем неполные строки. Также необходимо указать регион относительно чекбокса, в котором должен находиться артикул. На выходе получаем таблицу, во втором столбце которой список выбранных артикулов.



(Рис. 87 Локатор «Таблица альтернатив»)

Тип группировки. Указывает по какому принципу будет выполняться группировка альтернатив в рамках столбцов таблицы. Варианты: без группировки, простая группировка, расширенная группировка.

- **Без группировки**. В данном случае последовательность ячеек в столбцах итоговой таблицы не изменяется, т.е. та же, что у альтернатив наследуемых подполей.
- **Простая группировка**. Если выбран этот вариант группировки, то ячейки из всех столбцов кроме первого, сравниваются геометрически: в данном случае вычисляется расстояние между центрами ячеек из ключевого (первого) столбца и проверяемого столбца, и как только ячейка из не ключевого столбца попадается достаточно близко к ячейке из ключевого столбца, то она (ячейка из не ключевого столбца) помещается в ту же строку что и близлежащая ключевая ячейка. Достаточно близко в данном случае – это расстояние менее наибольшей ширины из ячейки ключевого и не ключевого столбцов (по горизонтали) и высоты по вертикали.
- **Расширенная группировка**. В этом случае, ячейки из не ключевых столбцов перемещаются в строку с ячейкой из ключевого столба. Если ячейка из не ключевого столбца попала в область, рассчитанную относительно ячейки из ключевого столбца и указанных параметров для расчета области в не ключевом столбце.

Пересортировывать ключ. Если выбрана данная опция, то ячейки из ключевого столбца будут отсортированы в порядке уменьшения степени доверия, иначе последовательность будет та же что и в наследуемом подполе.

Удалять неполные строки. Если выбрана эта опция, то строки таблицы, которые, имеют хотя бы одну пустую ячейку будут удалены из таблицы.

Удалять строки без ключа. Если выбрана эта опция, то строки таблицы, которые, имеют пустую ячейку в первом (ключевом) столбце, будут удалены.

Наследуемые подполя инструментов извлечения. Содержат список подполей с настройками, из которых формируется таблица. Первая запись определяет ключевой столбец таблицы. Имеющиеся параметры: подполе инструмента извлечения, максимальное количество результатов, минимальный процент доверия, настройки расчета области применения ключа (X, Y, ширина, высота), использовать несколько ключей.

Подполе инструмента извлечения. Указывает подполе инструмента извлечения, альтернативы из которого будут заполнять столбцы таблицы.

Максимальное количество результатов. Если это значение не 0, то из указанного подполя в таблицу попадут только указанное количество альтернатив.

Минимальный процент доверия. Минимальный процент доверия для альтернативы наследуемого подполя, для попадания в столбец таблицы. Диапазон значений: 0-100.

Настройки расчета области применения ключа. Описывают область на репрезентации ключа, попадая в которую ячейки не ключевых столбцов могут быть перенесены в строку с ключевой ячейкой при использовании расширенной группировки.

X – отступ от левой границы ключа выраженный в долях от ширины ключа. Диапазон значений: -25.00-25.00.

Y – отступ от верхней границы ключа выраженный в долях от высоты ключа. Диапазон значений: -25.00-25.00.

W – ширина области, выраженная в процентах от ширины ключа. Диапазон значений: 0.00-25.00.

H – высота области, выраженная в процентах от высоты ключа. Диапазон значений: 0.00-25.00.

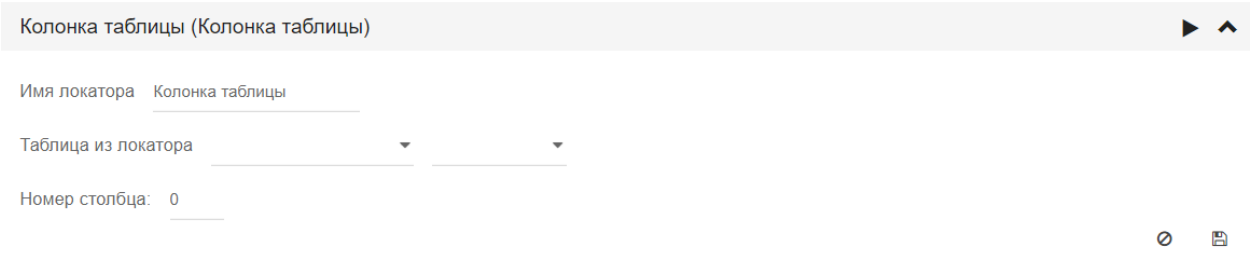
Несколько ключей. При выборе данной опции, при использовании расширенной группировки, одна и та же ячейка из не ключевого столбца может быть добавлена в строку с разными ключевыми ячейками, если она попадает сразу в несколько регионов.

Подполя инструмента извлечения:

Таблица. Подполе содержащее таблицу, сформированную из альтернатив наследуемых подполей.

17. Колонка таблицы ()

Берет из таблицы один столбец с указанным номером и записывает его ячейки в альтернативы своего подполя, так же может создавать альтернативы из строк таблицы.



(Рис. 88 Локатор «Колонка таблицы»)

Наследуемое табличное подполе. Указывает на подполе инструмента извлечения, табличного типа, столбец из которого будет использован в инструменте извлечения.

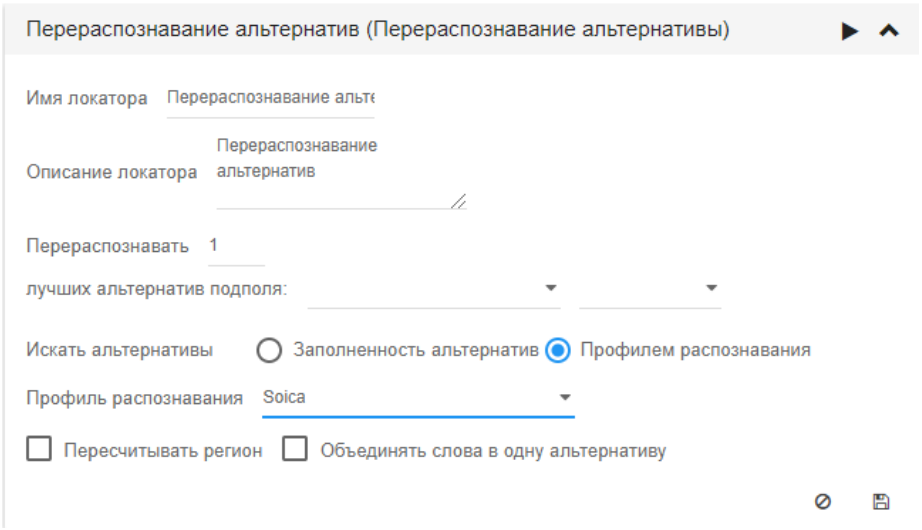
Номер извлекаемого столбца. Указывает номер столбца, который будет извлечен из таблицы и ячейки которого будут записаны в качестве альтернатив подполя инструмента извлечения, если значение не -1. Если же значение -1, тогда альтернативы инструмента извлечения формируются из строк таблицы.

Подполя инструмента извлечения:

Результат. Подполе, содержащее альтернативы, полученные из столбца наследуемой таблицы, либо из строк таблицы.

18. Перераспознование альтернатив ( )

Выполняет перечитывание указанного количества лучших (по степени доверия) альтернатив подполя заданным профилем распознавания. Есть опция менять размеры зоны по результатам распознанных результатов. Кроме того, полученные слова можно выводить как отдельные альтернативы или как одну альтернативу.



(Рис. 89 Локатор «Перераспозование альтернатив»)

Наследуемое подполе инструмента извлечения. Указывает подполе локатора, альтернативы которого будут перераспознаваться.

Количество перераспознаваемых альтернатив. Количество лучших (по степени доверия) альтернатив наследуемого подполя локатора, которые будут перераспознаны.

Заполненность альтернатив. При выборе данной опции включается проверка заполненности области альтернативы с вариантом упрощенного способа и с порогом заполненности.

Профиль распознавания. Указывает на профиль распознавания, которым будут перераспознаны области альтернатив наследуемого локатора.

Пересчитывать регион. Если данная опция включена, то размеры и координаты итоговых альтернатив могут быть скорректированы исходя из результатов распознавания, по объединенной области распознанных слов.

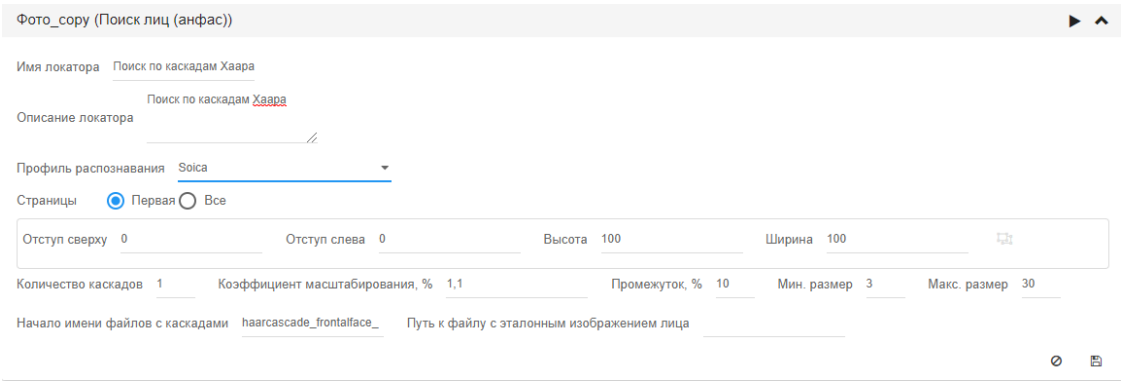
Объединять слова в одну альтернативу. Если данная опция включена, то все найденные при перераспознавании слова будут записаны в одну альтернативу и будут разделены пробелами, иначе из каждого слова будет создана своя альтернатива.

Подполя инструмента извлечения:

Результат. Подполе, содержащее альтернативы, полученные при перераспознавании альтернатив наследуемого подполя.

19. Поиск по каскадам Хаара ( )

Локатор позволяет производить поиск объекта с помощью каскадов Хаара (заранее настроенными признаками объекта на изображении). Указывается профиль распознавания и область поиска, а также на первой или на всех страницах будет осуществлен поиск. Так же указываются настройки поиска каскадом. Задается начало имени тренированного файла с каскадами и количество каскадов для поиска. В системе имеется 4 файла каскадов для поиска лица. Процент доверия вычисляется следующим образом: количество найденных объектов в одном месте/количество каскадов для поиска.



(Рис. 90 Локатор «Поиск по каскадам Хаара»)

Профиль распознавания. Указывает репрезентацию в которых будет осуществлен поиск объектов по указанным каскадам.

Настройки области инструмента извлечения. Описывает область репрезентации, в которой будет выполняться поиск объектов по указанным каскадам.

Отступ слева – отступ от левой границы изображения репрезентации выраженный в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.

Отступ сверху – отступ от верхней границы изображения репрезентации выраженный в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.

Ширина – ширина области, выраженная в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.

Высота – высота области, выраженная в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.

Суммы Отступ слева, Ширина и Отступ сверху, Высота так же не могут быть больше 100.

Страницы, на которых будет выполнятся инструмент извлечения. Возможные варианты – первая страница, все страницы. При поиске на **первой странице**, будет осуществляться поиск объектов по указанным каскадам в репрезентации указанного профиля только первой страницы документа. При поиске на **всех страницах**, будет осуществляться поиск объектов по указанным каскадам в репрезентации указанного профиля на каждой из страниц документа.

Количество каскадов. Максимальное количество файлов с каскадами, по которым будет производится поиск объектов на репрезентации. Диапазон значений: 1-4.

Фактор масштаба. Параметр, показывающий на сколько будет меняться масштаб изображения, каждый проход поиска. Чем он меньше, тем дольше и подробнее будет выполняться поиск. Диапазон значений: 0,5-3.

Промежуток. Параметр, определяющий, сколько соседей должен иметь каждый прямоугольник-кандидат. Этот параметр влияет на качество обнаруженных лиц. Более высокое значение приводит к меньшему количеству обнаружений, но с более высоким качеством.

Минимальный размер. Указывает минимально возможный размер обнаружаемого объекта. Параметр выражается в процентах от ширины и высоты изображения. Диапазон значений: 1-100.

Максимальный размер. Указывает максимально возможный размер обнаружаемого объекта. Параметр выражается в процентах от ширины и высоты изображения. Диапазон значений: 1-100.

Начало имени файлов с каскадами. Сами каскады хранятся в виде xml файлов, для поиска лиц их 4 шт. Имена этих файлов начинаются одинаково. Выполняется перебор xml файлов в корневом каталоге системы, и если файл начинается с «Начала имени файлов каскадов», то выполняется поиск объектов по каскадам из этого файла.

Путь к файлу с эталонным изображением лица. При добавлении пути к эталонному изображению будет проходить сравнивание найденного лица с лицом на эталонном изображении. В альтернативы записывается процент похожести.

Подполя инструмента извлечения:

Результат. Подполе, содержащее альтернативы, полученные в процессе поиска объектов по каскадам Хаара, с текстом True.

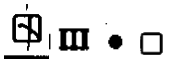
20. Поиск чекбоксов ( )

Служит для поиска чекбоксов – квадратов для заполнения. В настройках нужно указать репрезентацию, выбрать первую или все страницы, а также указать область поиска. Для лучшего результата репрезентация не должна быть монохромной, чтобы метка

заполнения либо не пересекала границы квадрата, либо отличалась по яркости. В настройках поиска указываются параметры предобработки изображения, пороги для бинаризации для определения границ и заполненности чекбокса. Также можно включить простой способ поиска чекбоксов, при котором не выполняется поиск внешней и внутренней границы рамки, но при этом в результаты попадает много лишнего.

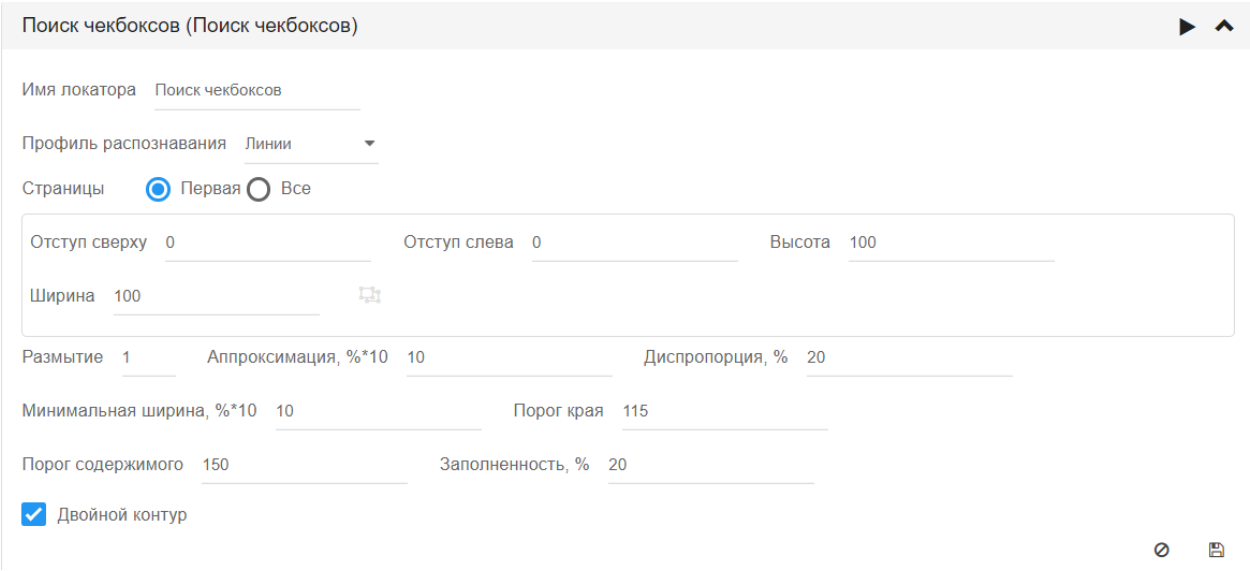
 - положительные чекбоксы

 - отрицательные чекбоксы

 - положительные чекбоксы, при бинарном изображении, с простым поиском.

 - отрицательные чекбоксы, при бинарном изображении, с простым поиском.

 - положительные чекбоксы, при бинарном изображении с расширенным поиском.



(Рис. 91 Локатор «Поиск чекбоксов»)

Профиль распознавания. Указывает репрезентацию в которых будет осуществлен поиск флажков.

Настройки области инструмента извлечения. Описывает область репрезентации, в которой будет выполняться поиск флажков.

Отступ слева – отступ от левой границы изображения репрезентации выраженный в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.

Отступ сверху – отступ от верхней границы изображения репрезентации выраженный в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.

Ширина – ширина области, выраженная в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.

Высота – высота области, выраженная в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.

Суммы Отступ слева, Ширина и Отступ сверху, Высота так же не могут быть больше 100.

Страницы, на которых будет выполнятся инструмент извлечения. Возможные варианты – первая страница, все страницы. При поиске на **первой странице**, будет осуществляться поиск флажков в репрезентации указанного профиля только первой страницы документа. При поиске на **всех страницах**, будет осуществляться поиск флажков в репрезентации указанного профиля на каждой из страниц документа.

Размытие. Указывает сколько раз будет выполняться размытие изображение. Служит для удаления шумов. Диапазон значений: 0-5.

Аппроксимация. Указывает коэффициент аппроксимации контуров для поиска рамок флажков, выраженный в десятых долях процента от длины линии контура. Диапазон значений: 0-1000.

Диспропорция. Указывает максимальное различие между высотой и шириной прямоугольника, при котором этот прямоугольник может считаться частью рамки флажка. Диапазон значений: 0-100.

Минимальная ширина. Указывает минимально возможную ширину прямоугольника, который может считаться частью рамки флажка, выраженный в десятых долях процента от ширины изображения. Диапазон значений: 0-1000.

Порог края. Указывает порог для бинаризации изображения результат которого будет использоваться для поиска рамок флажков. Диапазон значений: 0-255.

Порог содержимого. Указывает порог для бинаризации изображения флажка результат которого будет использоваться для расчета заполненности флажка. Диапазон значений: 0-255.

Заполненность. Величина, указывающая при достижении какого процента черных пикселей внутри рамки флажка после бинаризации флажок будет считаться выбранным. Диапазон значений: 0-100.

Двойной контур. При выборе этой опции рамка флажка будет определятся двумя вложенными прямоугольными контурами, иначе рамка флажка будет определяться одним прямоугольником.

Подполя инструмента извлечения:

Результат. Подполе, содержащее альтернативы, полученные в процессе поиска флажков, с текстом True, если флажок выбран или False, если флажок не выбран.

21. Выбор таблицы ()

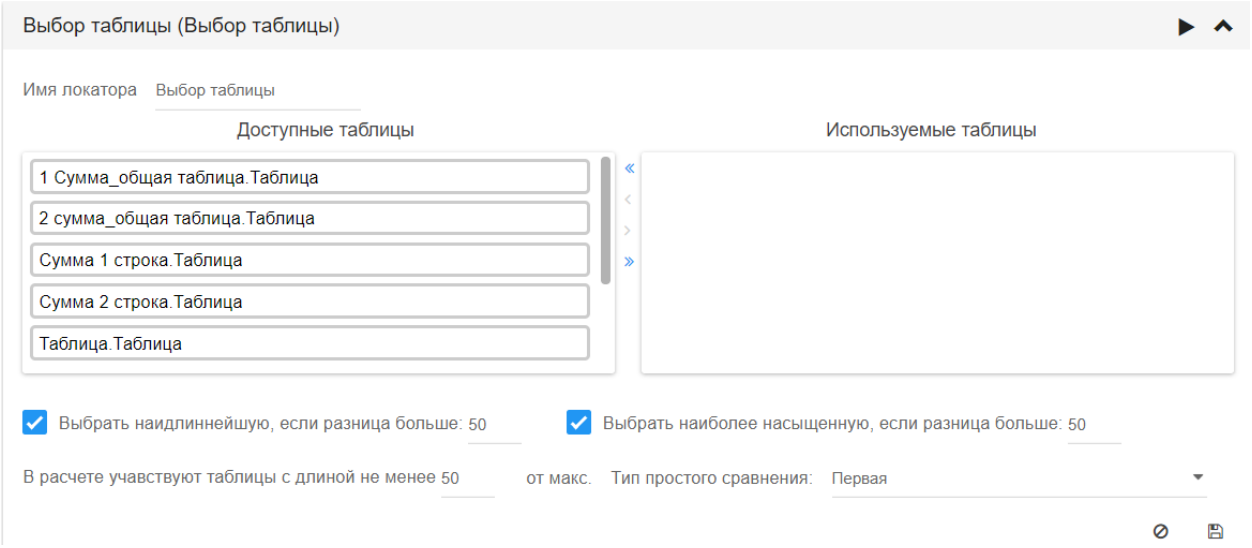
Служит для выбора одной лучшей таблицы из нескольких. Существуют следующие опциональные параметры выбора:

Выбрать самую длинную (по наибольшему количеству строк), если разница между количеством строк у самой длинной и второй по длине больше указанного значения в процентах от максимальной длины. Т.е. если эта опция выбрана, и порог установлен в 50%.

То при таблицах в 25 и 10 строк, в качестве лучшей выберется таблица из 25 строк и на этом инструмент извлечения завершит выполнение. Если же у второй по величине таблицы будет 15 строк, то это условие не сработает и будет рассчитываться следующий этап.

Выбрать самую насыщенную, если разница насыщенности между самой насыщенной и второй по насыщенности больше указанного числа. Под насыщенностью понимается среднее количество символов на ячейку. Принцип тот же что и в предыдущем случае, но при этом на участвующие в сравнении таблицы можно наложить отдельно еще и ограничение по минимальной длине в % от максимальной длины таблицы. Это нужно для того чтобы отсеять таблицы, например, с одной ячейкой и т.д.

Если же и это условие не прошло (или не было включено), тогда происходит выбор наилучшего без расчета разницы по следующим параметрам: количество строк, количество охваченных страниц, насыщенность, усредненная степень доверия, либо первая непустая в списке.



(Рис. 92 Локатор «Выбор таблицы»)

Наследуемые таблицы. Подполя, содержащие таблицы, из которых будет выбираться наилучшая по указанным параметрам.

Выбрать наидлиннейшую таблицу. Если выбрана данная опция, то в случае если самая длинная (по количеству строк) таблица больше второй по длине таблицы на указанное значение, то она (наидлиннейшая) выбирается в качестве результирующей.

Величина разницы по длине. Процент от количества строк наидлиннейшей (по количеству строк) таблицы. Если вторая по длине таблица меньше первой на указанный процент, и если выбрана опция «выбрать наидлиннейшую таблицу», то наидлиннейшая таблица будет выбрана в результат. Диапазон значений: 0-100.

Выбрать наиболее насыщенную таблицу. Если выбрана данная опция, то в случае если самая насыщенная (по среднему количеству символов в ячейке) таблица больше второй по насыщенности таблицы на указанное значение, то она (наиболее насыщенная) выбирается в качестве результирующей.

Величина разницы по насыщенности. Процент от количества строк наиболее насыщенной (по среднему количеству символов в ячейке) таблицы. Если вторая по

насыщенности таблица меньше первой на указанный процент, и если выбрана опция «выбрать наиболее насыщенную таблицу», то наиболее насыщенная таблица будет выбрана в результат. Диапазон значений: 0-100.

Величина разницы по длине при использовании наиболее насыщенной таблицы. Показывает процент от длины (по количеству строк) от наидлиннейшей таблицы, меньше которого таблицы не будут участвовать в сравнении по насыщенности (по среднему количеству символов в ячейке). Диапазон значений: 0-100.

Тип простого сравнения. В случае, если не выбрана или не выполнилась не одна из опций из «Выбрать наидлиннейшую таблицу» и «Выбрать наиболее насыщенную таблицу», то итоговая таблица будет выбрана исходя из одного из правил: наидлиннейшая, наиболее насыщенная, охватывающая наибольшее количество страниц, с наибольшей степенью доверия, первая.

Наидлиннейшая. В качестве результата выбирается таблица с наибольшим количеством строк.

Наиболее насыщенная. В качестве результата выбирается таблица с набольшим средним количеством символов на ячейку.

Охватывающая наибольшее количество страниц. В качестве результата выбирается таблица, которая располагается на большем количестве страниц в документе.

С наибольшей степенью доверия. В качестве результата выбирается таблица, у которой средняя степень доверия ячейки наибольшая.

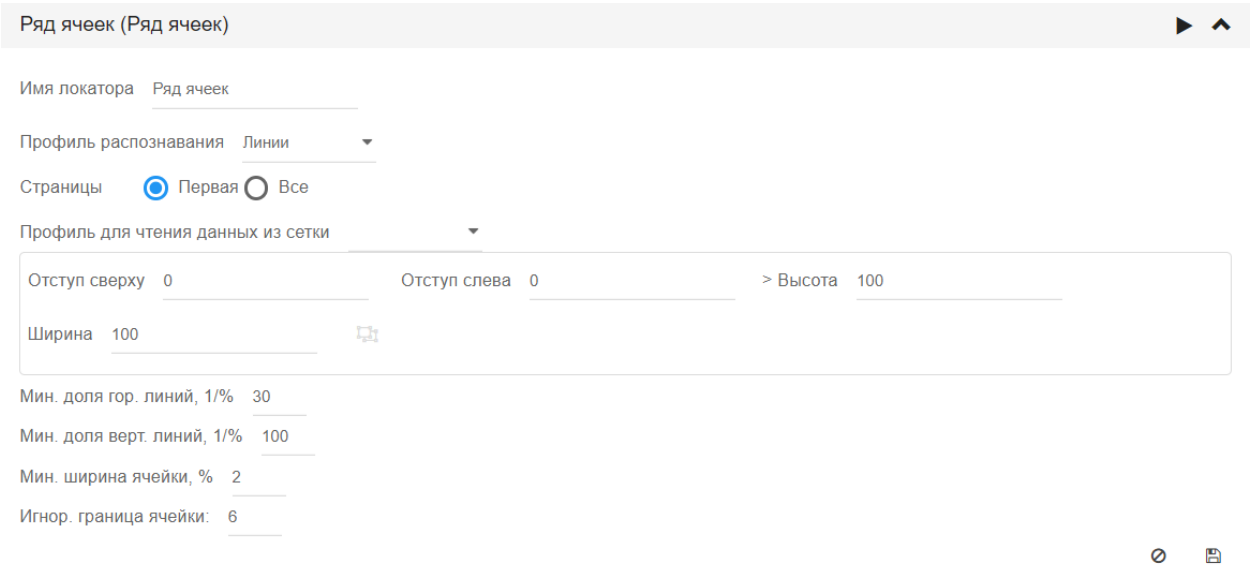
Первая. В данном случае выбирается первая из списка не пустая таблица.

Подполя инструмента извлечения:

Таблица. Подполе содержащее таблицу, выбранную из альтернатив наследуемых подполей.

22. Ряд ячеек ( )

Служит для поиска и распознавания анкетных сеток. Выполняет поиск линий способом, затем объединяет их в своего рода таблицы и выполняет перечитывание каждой ячейки указанным профилем. Поиск осуществляется в указанном профиле, страницеи заданной области. Создает альтернативу в подполе по каждой найденной сетке.



(Рис. 93 Локатор «Ряд ячеек»)

Профиль распознавания. Указывает репрезентацию, в которой будет осуществлен поиск анкетных сеток.

Настройки области инструмента извлечения. Описывает область репрезентации, в которой будет выполняться поиск анкетных сеток.

Отступ слева – отступ от левой границы изображения репрезентации выраженный в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.

Отступ сверху – отступ от верхней границы изображения репрезентации выраженный в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.

Ширина – ширина области, выраженная в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.

Высота – высота области, выраженная в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.

Суммы Отступ слева, Ширина и Отступ сверху, Высота так же не могут быть больше 100.

Страницы, на которых будет выполнятся инструмент извлечения. Возможные варианты – первая страница, все страницы. При поиске на **первой странице**, будет осуществляться поиск анкетных сеток в репрезентации указанного профиля только первой страницы документа. При поиске на **всех страницах**, будет осуществляться поиск анкетных сеток в репрезентации указанного профиля на каждой из страниц документа.

Профиль перераспознавания. Указывает профиль распознавания, которым будут перераспознаны ячейки из найденных анкетных сеток.

Минимальная доля горизонтальной линии. Указывает минимальную длину находимой горизонтальной линии для построения сетки, выраженную в обратных долях от ширины изображения. Т.е. 3 – это 1/3 ширины. Диапазон значений: 1-150.

Минимальная доля вертикальной линии. Указывает минимальную длину находимой вертикальной линии, выраженную в обратных долях от ширины изображения. Т.е. 3 – это 1/3 ширины. Диапазон значений: 1-150.

Минимальная ширина ячейки. Указывает минимально возможную ширину ячейки в анкетной сетке, выраженную в процентах от ширины изображения. Диапазон значений: 1-100.

Игнорируемая граница ячейки. Указывает отступ от границы ячейки внутрь для формирования области перераспознавания, выраженный в пикселях.

Подполя инструмента извлечения:

Результат. Подполе, содержащее альтернативы, полученные в процессе поиска анкетных сеток и их перераспознавания.

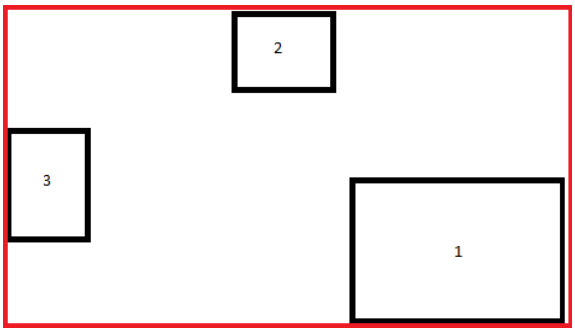
Сетки – наборы ячеек в которых был произведено перераспознование.

Линии – найденные линии на документе, которые затем собираются в сетки

23. Извлечение по зонам ( )

Служит для выделения областей в указанной репрезентации в качестве альтернатив подполей, и заполнения их из результатов ocr. Для одного локатора можно задать до 15ти зон.

Алгоритмы расчета зоны различаются по количеству используемых ключевых зон. Если ключевых зон нет, то зоны рассчитываются по следующему принципу:



Минимальный X всех зон сопоставляет с левой границей изображения, минимальный Y – с верхней, максимальный X – с правой, максимальный Y – нижней. Размеры зон изменяются пропорционально указанным значениям. Например, размер изображения 100*200, есть 2 зоны: x=50, y=0, w=20, h=10 и x=15, y=15, w=70, h=45. Минимальный X для зон – 15, максимальный – 85, минимальный Y – 0, максимальный – 60. В данном случае получается множитель для зоны по X = 100/(85-15) = 1.42, а для Y = 200/(60-0) = 3,33. Соответственно смещая зоны к левому верхнему углу изображения получаем итоговые координаты и размеры этих 2х зон:

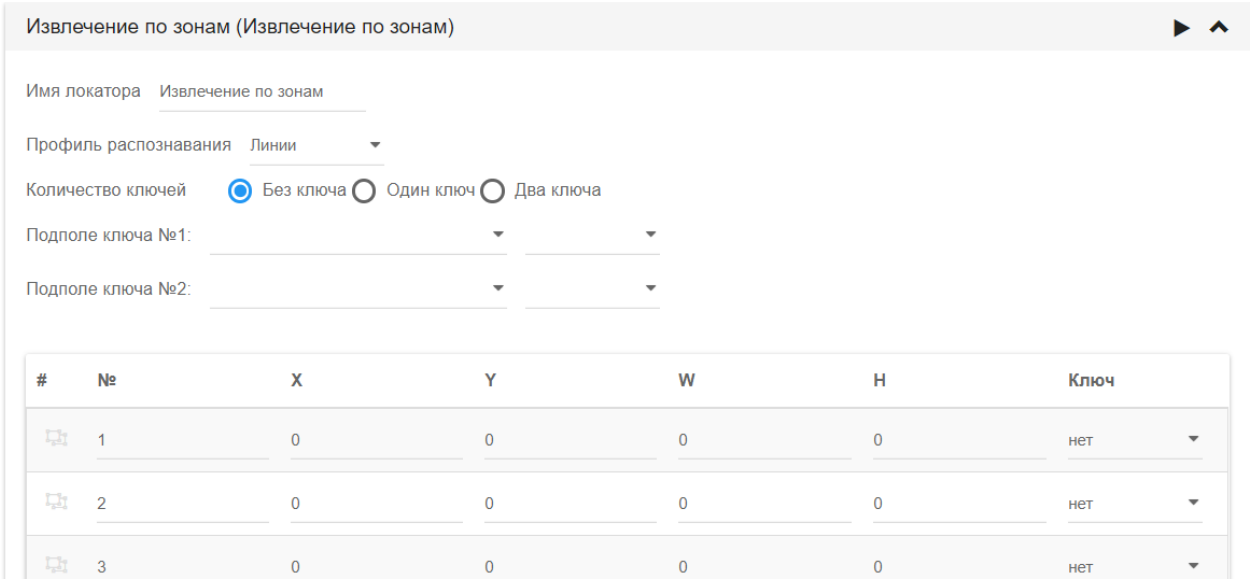
X = (50-15)*1.42=49.7, Y = 0 * 3.33 = 0, W=20*1.42=28.4, H=10*3.33=33.3;

X = (15-15)*1.42=0, Y = 15 * 3.33 = 49.95, W=20*1.42=99.4, H=45*3.33=149,85.

В случае использования одного ключа, пропорции вычисляются исходя из соотношений реальных размеров альтернатив и указанных настроек зоны для ключа. В случае же использовании двух ключей пропорции еще зависят от взаимного расположения этих ключей в оригинале и в настройках зон.

По сути инструмент извлечения зоны с одним ключом это 14 Position инструментов извлечения, с возможностью более тонкой настройки зоны.

В результате инструмента извлечения 15 подполей.



(Рис. 93 Локатор «Извлечение по зонам»)

Профиль распознавания. Указывает репрезентацию, изображение которой будет делиться на зоны, и из ocr которой они будут наполняться.

Количество ключей. Указывает количество ключевых альтернатив подполей, относительно которых будет производиться расчет зон. Варианты: без ключей, один ключ, два ключа.

- **Без ключей**. В данном случае расчет размеров и координат зон будет производится относительно размеров изображения репрезентации.
- **Один ключ**. В данном случае расчет размеров и координат зон будет производится относительно размеров и координат альтернативы указанного ключа.
- **Два ключа**. В данном случае расчет размеров и координат зон будет производится относительно размеров, координат и взаимного расположения альтернатив указанных ключей.

Подполе первого ключа. Указывает на подполе инструмента извлечения лучшая (по степени доверия) альтернатива которого, будет использоваться в качестве первого ключа.

Подполе второго ключа. Указывает на подполе инструмента извлечения лучшая (по степени доверия) альтернатива которого, будет использоваться в качестве второго ключа.

Список зон. Содержит список с настройками для определения зон. Настройки: номер зоны, регион зоны (X, Y, ширина, высота), ключ.

Номер зоны. Номер зоны в результирующим подполе.

Регион зоны. Указывает область зоны в относительных единицах (X, Y, ширина, высота).

Ключ. Указывает, является ли это поле ключевым для первого или второго ключа.

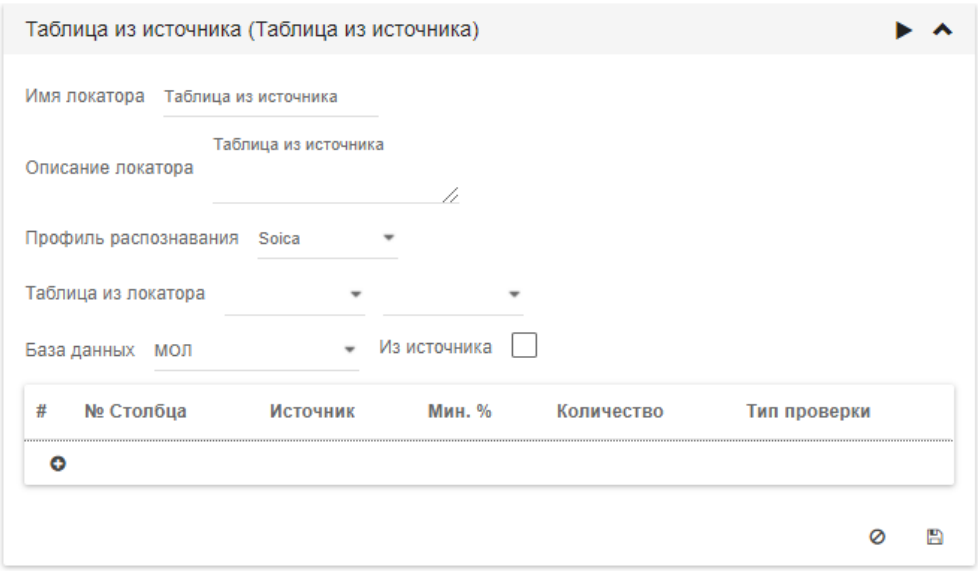
Подполя инструмента извлечения:

Подполе зоны (15 штук). Подполе, содержащее альтернативы, полученные в процессе расчета координат и размеров зон и заполненные результатами ocr из укаанной репрезентации.

24. Таблица из источника ( )

Предназначен для наполнения таблицы из источника данных, из источника данных берутся только те строки что удовлетворяют условиям. В условиях указывается подполе, из которого берется указанное количество лучших альтернатив, с процентом доверия превышающим указанное, номер столбца источника данных с которым будет выполняться сравнение альтернатив и тип сравнения: Равен, Содержит, Содержится.

Например, чтобы получить всех поставщиков из города заказчика, необходимо выбрать справочник поставщиков, столбец с адресом, тип сравнения «содержит». А в качестве подполя указать подполе с найденным наименованием города заказчика.



(Рис. 94 Локатор «Таблица из источника»)

Профиль распознавания. Указывает репрезентацию, на которой будет отображаться сформированная таблицы.

В данном локаторе можно наследовать Таблицу из локатора или Базу данных. Для этого нужно выбрать нужную опцию.

Таблица из локатора. Табличное подполе другого локатора. Правила, применимые к указанному локатору, наследуются.

База данных. Ссылка на таблицу с данными, ячейки строк которой будут сопоставляться с альтернативами из указанных подполей.

Из источника. При включенной опции при выполнении локатора таблица с данными будет наполняться из файла или подключенной таблицы из базы данных. Иначе таблица будет наполняться из временной копии, сохраненной в базе данных системы.

Список настроек извлечения строк из БД:

- **Номер столбца в БД**. Номер столбца в таблице из указанного источника данных с ячейками которого будет производится сравнение альтернатив указанного подполя.
- **Источник**. Указывает на альтернативы, которые будут сравниваться с ячейками из указанного столбца источника данных.
- **Минимальный процент доверия**. Указывает минимальный процент доверия альтернатив наследуемого подполя инструмента извлечения которые будет сравниваться с ячейками из источника данных. Диапазон значений: 0-100.
- **Количество**. Указывает максимальное количество лучших (по степени доверия) альтернатив которые будут сравниваться с ячейками из источника данных.
- **Тип проверки**. Принцип сравнения альтернатив подполя. Варианты: равен, содержится, содержит.

— **Равен**. Строки из источника данных выбираются в том случае, если альтернатива инструмента извлечения совпадает по тексту с ячейкой из указанного столбца источника.

— **Содержит**. Строки из источника данных выбираются в том случае, если альтернатива инструмента извлечения содержит текст из ячейки из указанного столбца источника.

— **Содержится**. Строки из источника данных выбираются в том случае, если альтернатива инструмента извлечения содержится в тексте с ячейкой из указанного столбца источника.

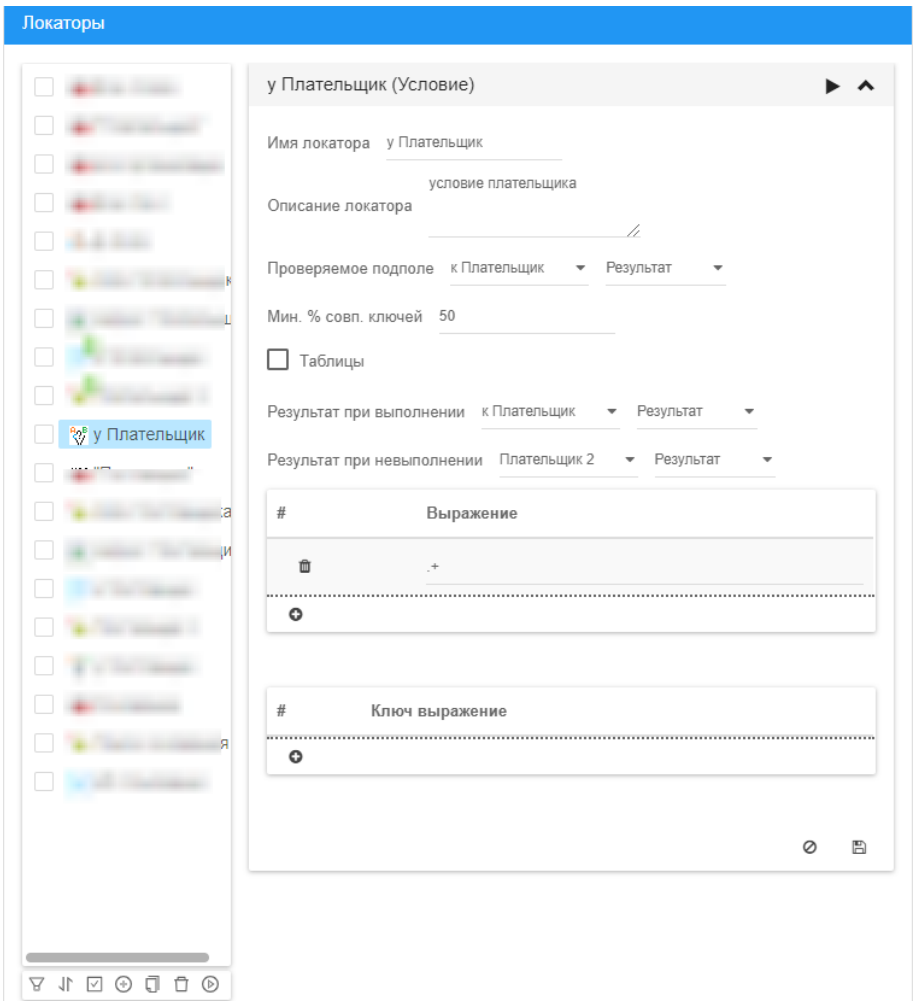
Подполя инструмента извлечения:

Таблица. Подполе содержащее таблицу, сформированную из выбранных строк источника данных.

25. Условие ( )

Локатор проверяет один из выполненных локаторов на заданное условие и в зависимости от выполнения условия, либо не выполнения выбирает себе альтернативы указанных локаторов. В качестве условия может быть проверка регулярным выражением или совпадение с ключом больше указанного процента. Локатор может выбирать между подполями с альтернативами или с таблицами.

В случае если альтернатив у проверяемого локатора нет, то выбирается подполе «Иначе».



(Рис. 95 Локатор «Условие»)

Проверяемое подполе. Подполе, лучшая (по степени доверия) альтернатива, которого проверяется на совпадение указанному ключу или регулярному выражению.

Минимальное совпадение ключей. Минимальный процент совпадения ключа и альтернативы проверяемого подполя, для того чтобы считать условие выполненным. Диапазон значений: 0-100.

Таблицы. При выборе этой опции, подполя при выполнении и невыполнении условия должны содержать таблицы.

Результат при выполнении. Указывает подполе локатора, альтернативы (либо таблицы) которого скопируются в альтернативу текущего локатора при выполнении условия.

Результат при невыполнении. Указывает подполе локатора, альтернативы (либо таблицы) которого скопируются в альтернативу текущего локатора при невыполнении условия.

Регулярное выражение. Регулярное выражение, соответствие с котором проверяется у альтернативы проверяемого подполя. Если регулярное выражение не указано, то происходит проверка на соответствие ключу.

Ключ. Текст, с которым сравнивается альтернатива проверяемого подполя, если не указано регулярное выражение.

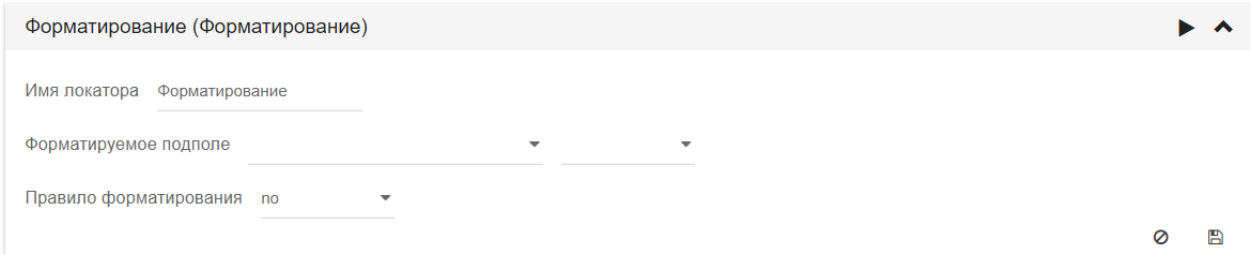
Подполя инструмента извлечения:

Результат. Подполе, содержащее альтернативы подполя, выбранного из наследуемых подполей.

Таблица. Подполе, содержащее таблицы из подполя, выбранного из наследуемых подполей табличного типа при выбранной опции «Таблица».

26. Форматирование ( )

Локатор применяет правила форматирования к альтернативам указанного подполя и записывает результат в качестве альтернатив своего подполя.



(Рис. 96 Локатор «Форматирование»)

Наследуемое подполе. Подполе локатора, альтернативы которого после форматирования, попадут в результирующее подполе.

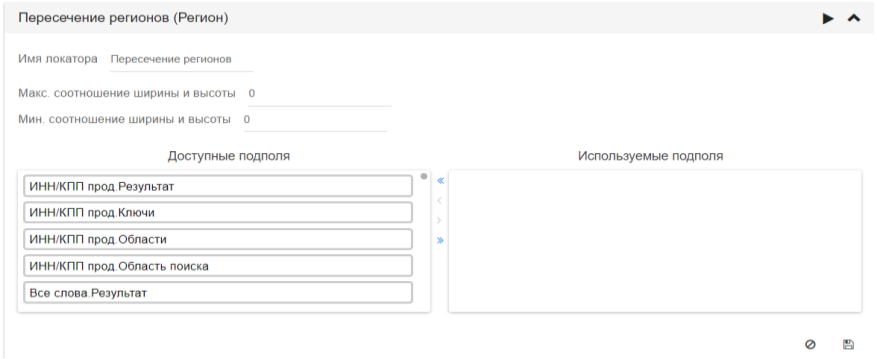
Правило форматирования. Указывает заранее настроенное правило форматирования, которое будет применено к альтернативам наследуемого подполя локатора.

Подполя инструмента извлечения:

Результат. Подполе, содержащее альтернативы наследуемого подполя после применения к ним указанного правила форматирования.

27. Пересечение регионов ( )

Локатор получает пересечения областей альтернатив наследуемых локаторов. Если все области с группы альтернатив пересекаются – на месте пересечения создается альтернатива. Максимальное количество альтернатив в результате – минимальное количество альтернатив в группах. Входящие альтернативы можно отсеивать по пропорциям.



(Рис. 97 Локатор «Пересечение регионов»)

Максимальное соотношение Ширина/Высота. Если значение этого параметра не 0, то альтернативы, у которых указанное соотношение больше указанного не участвуют в сравнении.

Минимальное соотношение Ширина/Высота. Если значение этого параметра не 0, то альтернативы, у которых указанное соотношение меньше указанного не участвуют в сравнении.

Доступные подполя – подполя локаторов, которые можно использовать для нахождения пересечения областей.

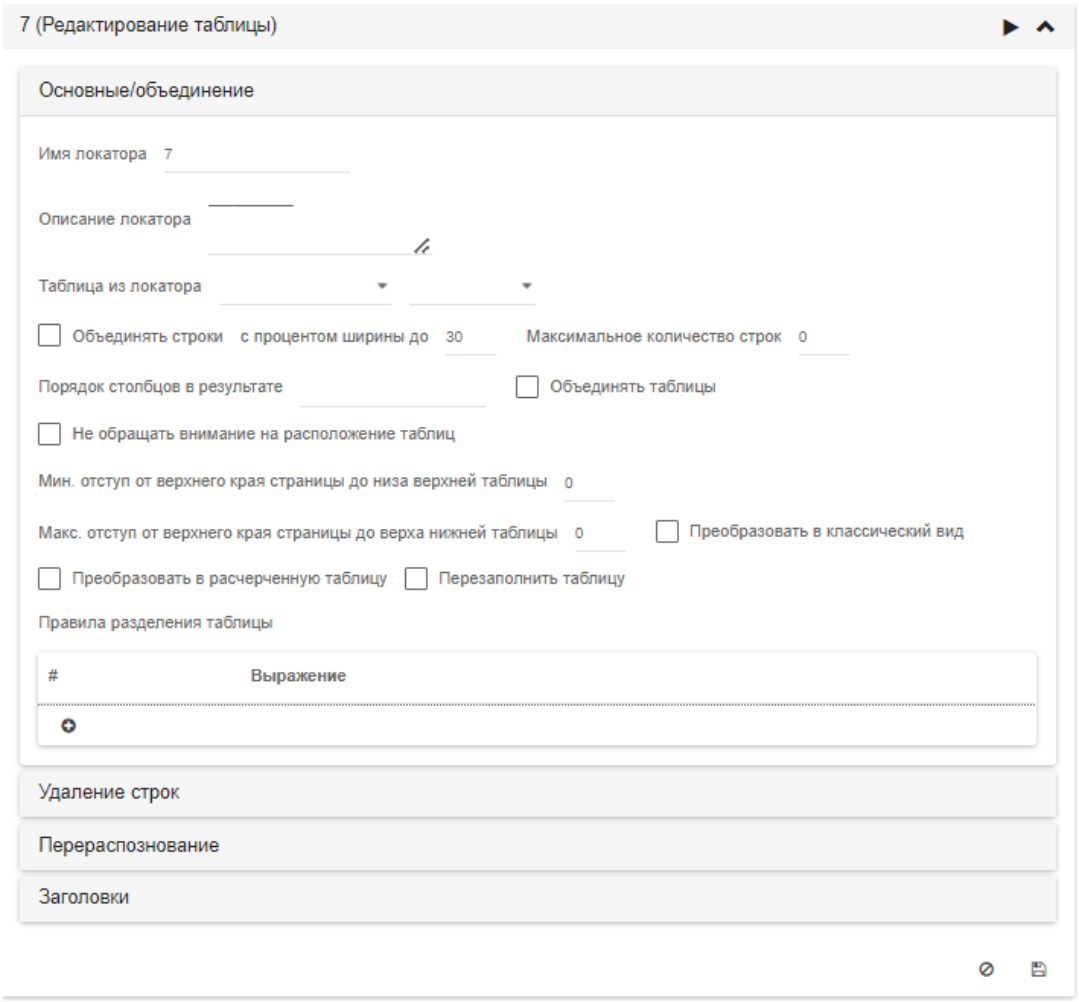
Используемые подполя. Подполя локаторов, место пересечения областей которых, если таковое имеется, будет записано в качестве альтернативы результирующего подполя.

Подполя инструмента извлечения:

Результат. Альтернативы, полученные в результате пересечения альтернатив с нескольких наследуемых подполей локатора. Размер альтернативы определяется областью в которой пересекаются области каждой альтернативы и наследуемых подполей. Если таковой области нет, то для этого ряда наследуемых альтернатив не создается результирующая альтернатива. Расчет пересечения выполняется только для полных рядов альтернатив наследуемых подполей.

28. Редактирование таблиц ()

Предназначен для корректировки итоговых альтернатив уже найденной таблицы. Локатор обладает частью функционала табличного локатора по постобработке таблиц, а именно: удаление строк, перераспознавание, заголовки, объединение строк, объединение таблиц с разных страниц. В данном случае этот функционал можно применить к таблице (таблицам), возвращаемой любым подполем подобного типа.



(Рис. 98 Локатор «Редактирование таблиц»)

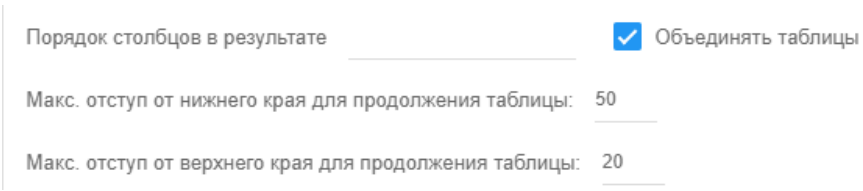
Таблица из локатора. Подполе локатора, содержащее таблицы. Которые после указанных преобразований, запишутся в результирующее подполе.

Объединять строки с процентом ширины до. При использовании объединения строк, количество столбцов в таблице умноженное на данное значение меньше чем количество заполненных ячеек, тогда эта строка объединяется с предыдущей, если таковая имеется. Диапазон значений: 0-100.

Максимальное количество строк. Ограничение количества строк в итоговой таблице.

Порядок столбцов в результате. В поле указывается через запятую порядок столбцов, в котором они будут передаваться в результирующую альтернативу локатора.

Объединять таблицы. Объединение таблиц с разных страниц в одну в том случае, когда таблица начинается на одной странице, а заканчивается на другой. Необходимо указать Максимальные отступы от нижнего и от верхнего края для продолжения таблицы.



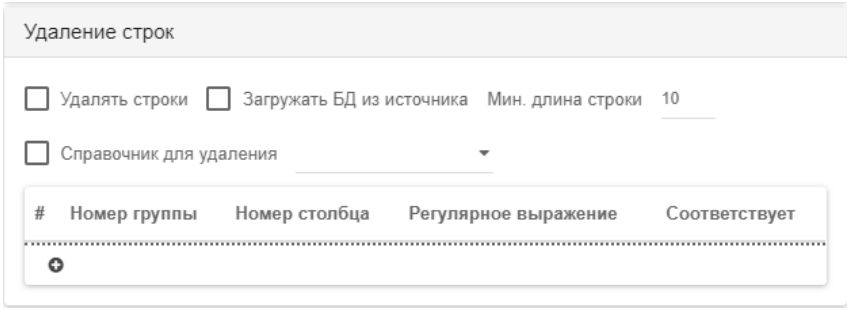
Не обращать внимание на расположение таблиц. Позволяет объединять таблицы не зависимо от расположения таблиц на многостраничном документе.

Преобразовать в классический вид. Преобразование таблицы с объединенными ячейками из локатора Блоки SOICAII в «классический» матричный вид.

Преобразовать в расчерченную таблицу. Преобразование таблицы с объединенными ячейками из локатора Блоки SOICAII в расчерченную таблицу.

Перезаполнить таблицу. Позволяет перезаполнить вновь полученные ячейки текстом из OCR. Используется вместе с опцией «Преобразовать в расчерченную таблицу». OCR берется из того же профиля распознавания что и для преобразованной в расчерченную таблицу. Работает только для таблиц, полученных движком Soica.

Правила разделения таблицы. Позволяет разделить таблицу по регулярным выражениям на несколько таблиц.



Удалять строки. Если данная отметка стоит, то будет применяться удаление строк.

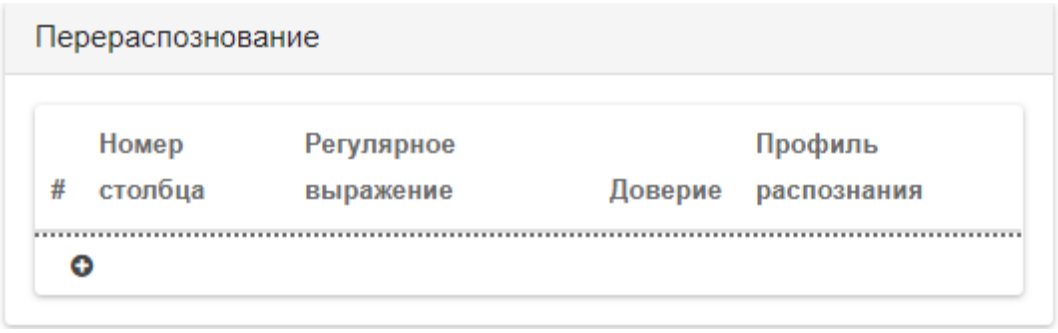
Загружать БД из источника. При включенной опции при выполнении локатора таблицы с данными (для удаляемых строк) будут наполняться из файла или подключенной таблицы из базы данных. Иначе таблицы будут наполняться из временной копии, сохраненной в базе данных системы.

Мин. длина строки. Указывает минимальное количество символов в строке таблицы при котором эта строка не будет удалена. Диапазон значений: 0-100.

Справочник для удаления. При использовании этой настройки, строки из указанного справочника будут сравниваться с текстом строк таблицы с помощью неточного соответствия, и в случае совпадения свыше 50% строка из таблицы будет удалена.

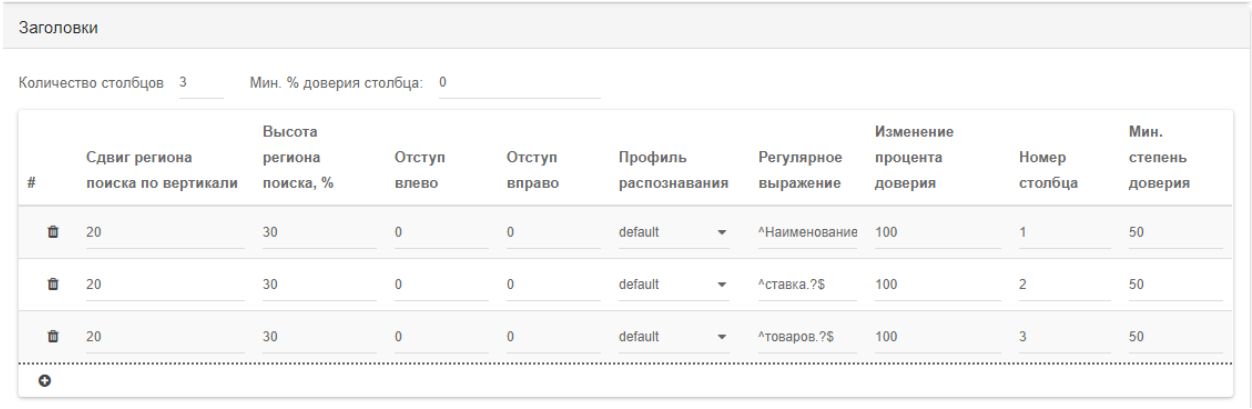
Правила удаления строк. Содержат коллекцию параметров для удаления строк из таблицы. Параметры: номер группы, номер столбца, регулярное выражение, соответствие.

- **Номер группы**. Определяет группы правил, которые выполняются, при выполнении всех условий в группе.
- **Номер столбца**. Определяет к какому столбцу будет применено указанное правило.
- **Регулярное выражение**. Указывает регулярное выражение для проверки соответствия ему ячеек из указанного столбца.
- **Соответствие**. При выбранной опции, строка будет удаляться если ячейка соответствует указанному регулярному выражению, при не выбранной опции – наоборот, если ячейка не соответствует регулярному выражению, произойдет удаление.



Настройки перераспознавания. Содержат коллекцию параметров, служащих для выполнения повторного распознавания областей ячеек таблицы при соответствующих условиях.

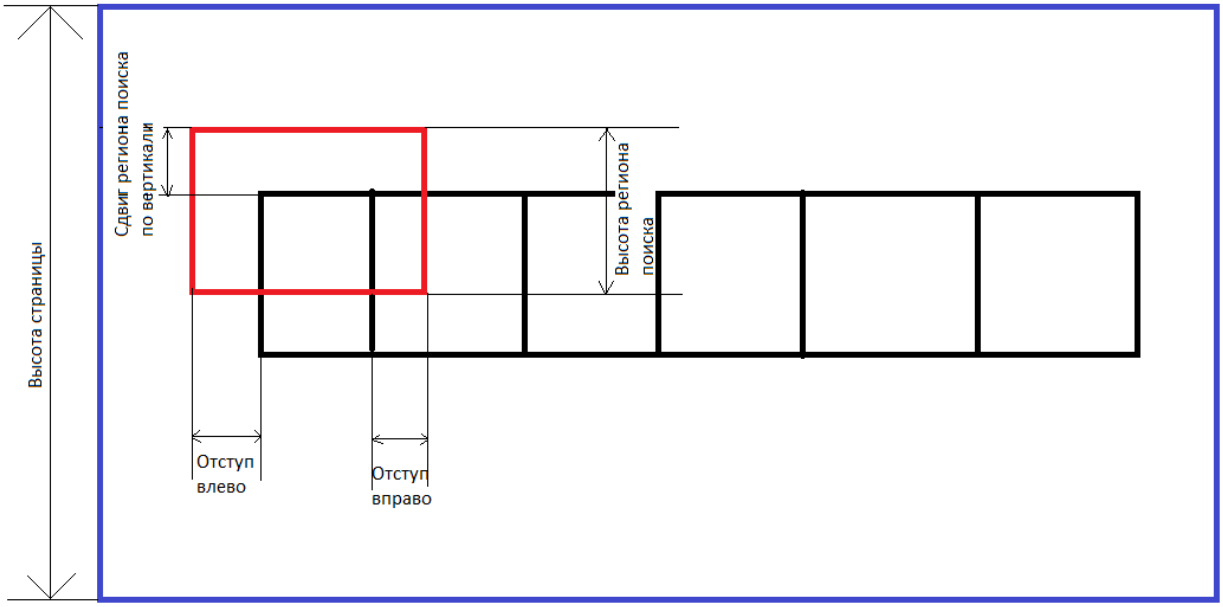
— **Номер столбца**. Указывается для какого столбца применяются настройки.
— **Регулярное выражение**. Указывает то регулярное выражение, при несоответствии текста ячейки которому, ячейка будет перераспознана.
— **Доверие**. Указывает минимальный процент доверия к ячейке таблицы указанного столбца при котором не будет требоваться ее перераспознавание. Диапазон значений: 0-100.
— **Профиль распознавания**. Указывает профиль распознавания, которым будет перераспознана ячейка, при выполнении соответствующих условий.



Количество столбцов. Количество столбцов в итоговой таблице при использовании заголовков.

Минимальный % доверия столбца. Процент доверия, при превышении результатом заголовков которого, столбец будет записан в итоговую таблицу. Диапазон значений: 0-100.

Коллекция правил заголовков. Содержит настройки для формирования результатов поиска заголовков. Указанные параметры: сдвиг региона по вертикали, высота региона поиска, регулярное выражение, изменение процента доверия, номер столбца, минимальная степень доверия.



Сдвиг региона поиска по вертикали. Указывает расстояние от верхней границы таблицы в процентах от высоты страницы на котором будет верхняя граница региона поиска признака столбца. Диапазон значений: -100 - 100.

Высота региона поиска, %. Указывает высоту региона поиска признака столбца, указанную в процентах от высоты страницы. Диапазон значений: 0 - 100.

Отступ влево. Указывает величину в процентах относительно ширины страницы с которой регион поиска, сдвинется влево от левой границы столбца.

Отступ вправо. Указывает величину в процентах относительно ширины страницы с которой регион поиска, сдвинется вправо от левой границы столбца.

Регулярное выражение. Указывает регулярное выражение, которому должно соответствовать слово в указанном регионе, для того чтобы записать его в результаты поиска заголовков.

Изменение процента доверия. Указывает величину, на которую меняется итоговая степень доверия классификации столбца. Диапазон значений: -100 - 100.

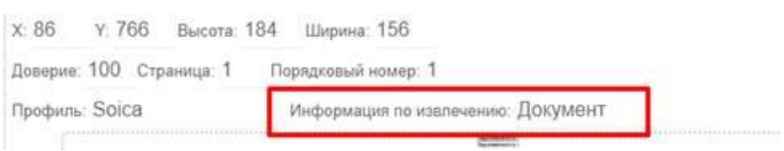
Номер столбца. Указывает на результат по какому столбцу будет влиять правило заголовка.

Минимальная степень доверия. Указывает минимальную степень доверия к слову из ocr, чтобы оно могло попасть в результаты поиска заголовков.

Подполя инструмента извлечения:

Таблицы. Наследуемая таблица, после выполненного редактирования указанными настройками.

Для удобства инженера реализован вывод регулярного выражения, по которому производится поиск ключа, у каждого региона:



А также отображение столбца, в котором найден ключ:

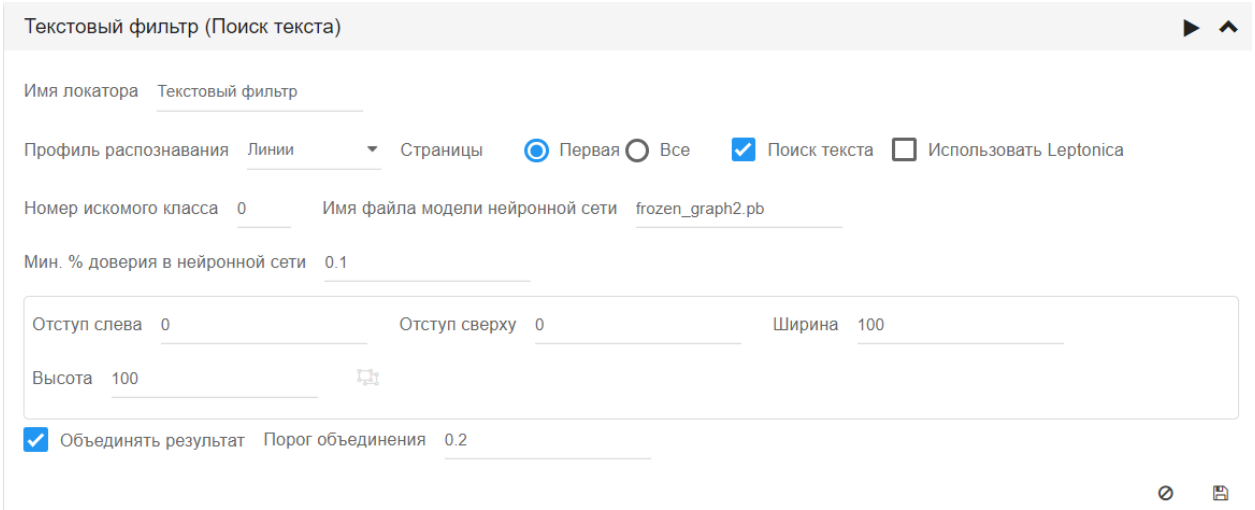
Первая цифра – номер столбца, который должен присвоится при нахождении ключа в регионе.

Вторая цифра – номер столбца основной таблицы, который проверяется на наличие ключа.

False\True – найден\не найден ключ в указанном регионе.

29. Текстовый фильтр ()

Локатор выполняет поиск текстовых областей на репрезентации с помощью нейросети и записывает найденные области в альтернативы.



(Рис. 99 Локатор «Текстовый фильтр»)

Профиль распознавания. Указывает репрезентацию в которых будет осуществлен поиск текстовых блоков.

Страницы, на которых будет выполнятся инструмент извлечения. Возможные варианты – первая страница, все страницы. При поиске на **первой странице**, будет осуществляться поиск текстовых блоков на изображении репрезентации указанного профиля только первой страницы документа. При поиске на **всех страницах**, будет осуществляться поиск текстовых блоков на изображении репрезентации указанного профиля на каждой из страниц документа.

Поиск текста. При поиске используется ocr репрезентации.

Использовать Leptonica. Используется нейросеть Leptonica.

Имя файла модели нейронной сети. Имя файла, который должен находиться в корневой папке приложения с обученной моделью для нейросети.

Минимальный процент доверия нейросети. Параметр указывающий минимальную степень доверия к результату поиска текстового блока, с которой этот блок будет учитываться. Диапазон значений: 0-100.

Настройки области инструмента извлечения. Описывает область репрезентации, в которой будет выполняться поиск текстовых блоков.

Отступ слева – отступ от левой границы изображения репрезентации выраженный в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.

Отступ сверху – отступ от верхней границы изображения репрезентации выраженный в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.

Ширина – ширина области, выраженная в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.

Высота – высота области, выраженная в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.

Суммы Отступ слева, Ширина и Отступ сверху, Высота так же не могут быть больше 100.

Объединять результаты. Опция, которая указывает на то, будут ли объединяться накладывающиеся текстовые блоки.

Порог объединения. Указывает процент пересечения текстовых блоков для их объединения. Диапазон значений: 0-100.

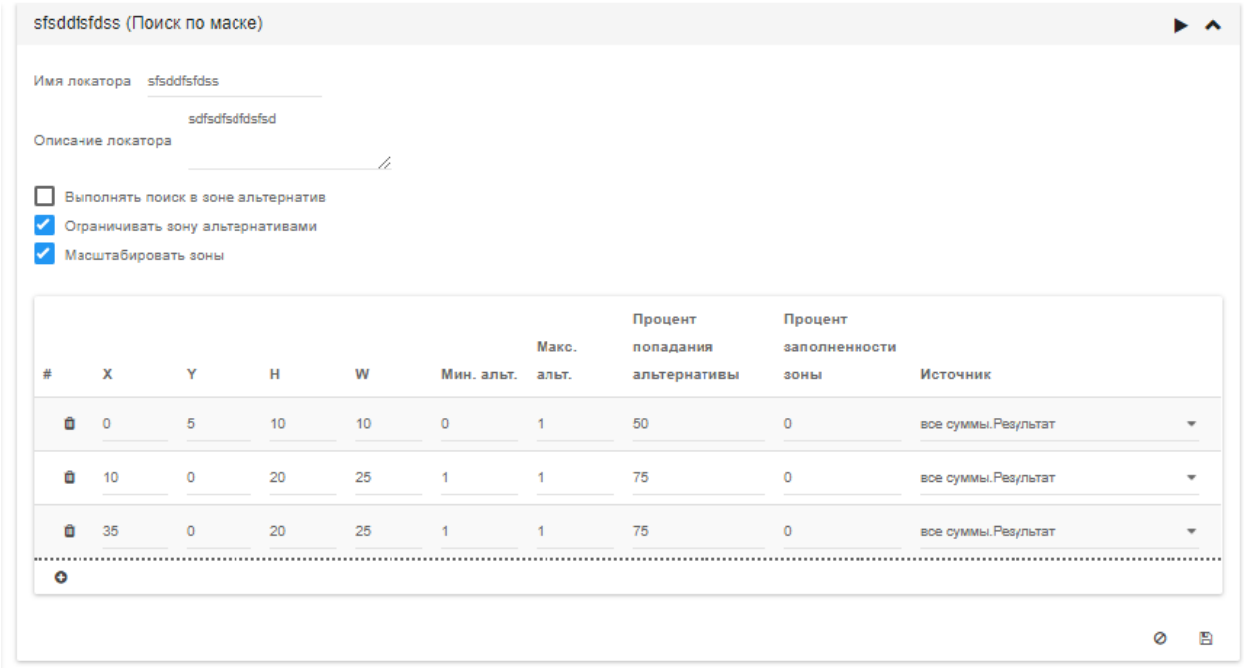
Подполя инструмента извлечения:

Результат. Содержит альтернативы с текстовыми блоками, текст отсутствует.

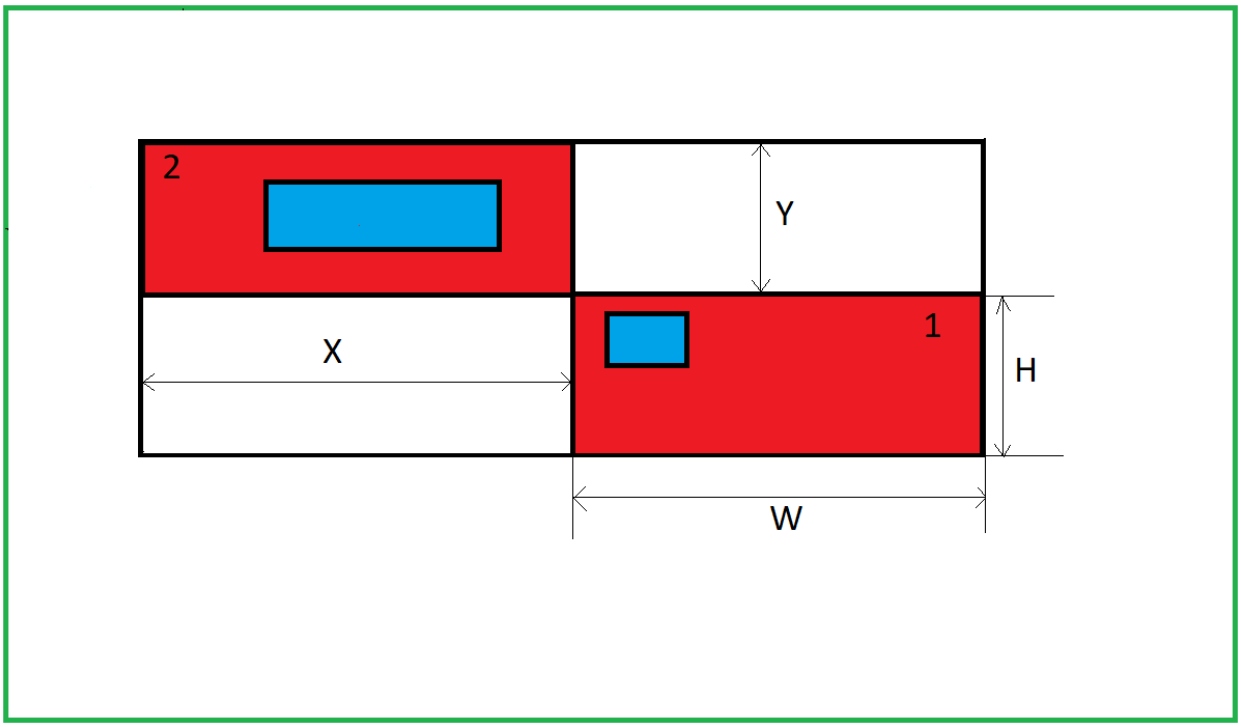
30. Поиск по маске ()

Локатор составляет таблицу из альтернатив наследуемого подполя.

Осуществляется поиск альтернатив, расположенных определенным образом относительно друг друга. Строки таблицы состоят из зон, составляющих маску локатора. Количество столбцов в таблице – количество зон в маске.



(Рис. 100 Локатор «Поиск по маске»)



При применении этого локатора создается графическая маска. Для каждой зоны локатора выбирается свой источник альтернатив.

Красным на рисунке выделены области поиска, зеленый прямоугольник - это границы страницы, синие прямоугольники внутри красных зон – это искомые альтернативы.

Наследуемое подполя. Подполе, альтернативы которого будут составлять строки итоговой таблицы, при соответствии их указанной маске.

Выполнять поиск в зоне альтернатив. Опция, которая указывает на то, относительно чего будут рассчитаны размеры и координаты зон маски, либо относительно размеров изображения первой репрезентации из альтернатив наследуемого подполя, либо исходя из размеров прямоугольника, созданного из всех альтернатив наследуемого подполя.

Ограничивать зону альтернативами. При выборе этой опции, размеры ячейки таблицы берутся не из пересчитанной зоны маски, а рассчитываются исходя из попавших в зону альтернатив.

Масштабировать зоны. При выборе этой опции, размер первой зоны будет изменен до размера рассматриваемой альтернативы. Размеры и положение остальных зон изменится пропорционально.

Список зон, составляющих маску. Включает в себя координаты и замер выраженный в процентах: X, Y, Ширина, Высота, Минимальное и максимальное количество альтернатив, попадающих в зону, Минимальный процент попадания альтернативы в зону, Минимальный процент заполненности зоны, источник.

Настройки области зоны. Описывает область зоны, составляющей маску.

- **X** – отступ от левой границы изображения репрезентации выраженный в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.
- **Y** – отступ от верхней границы изображения репрезентации выраженный в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.
- **W** – ширина области, выраженная в процентах от ширины изображения репрезентации. Диапазон значений: 0-100.
- **H** – высота области, выраженная в процентах от высоты изображения репрезентации. Диапазон значений: 0-100.

Минимальное количество альтернатив. Минимальное количество альтернатив, которое должно находиться в зоне (с процентом принадлежности зоне не ниже указанного). Диапазон значений: 0-100.

Максимальное количество альтернатив. Максимальное количество альтернатив, которое может находиться в зоне (с процентом принадлежности зоне не ниже указанного). Диапазон значений: 0-100.

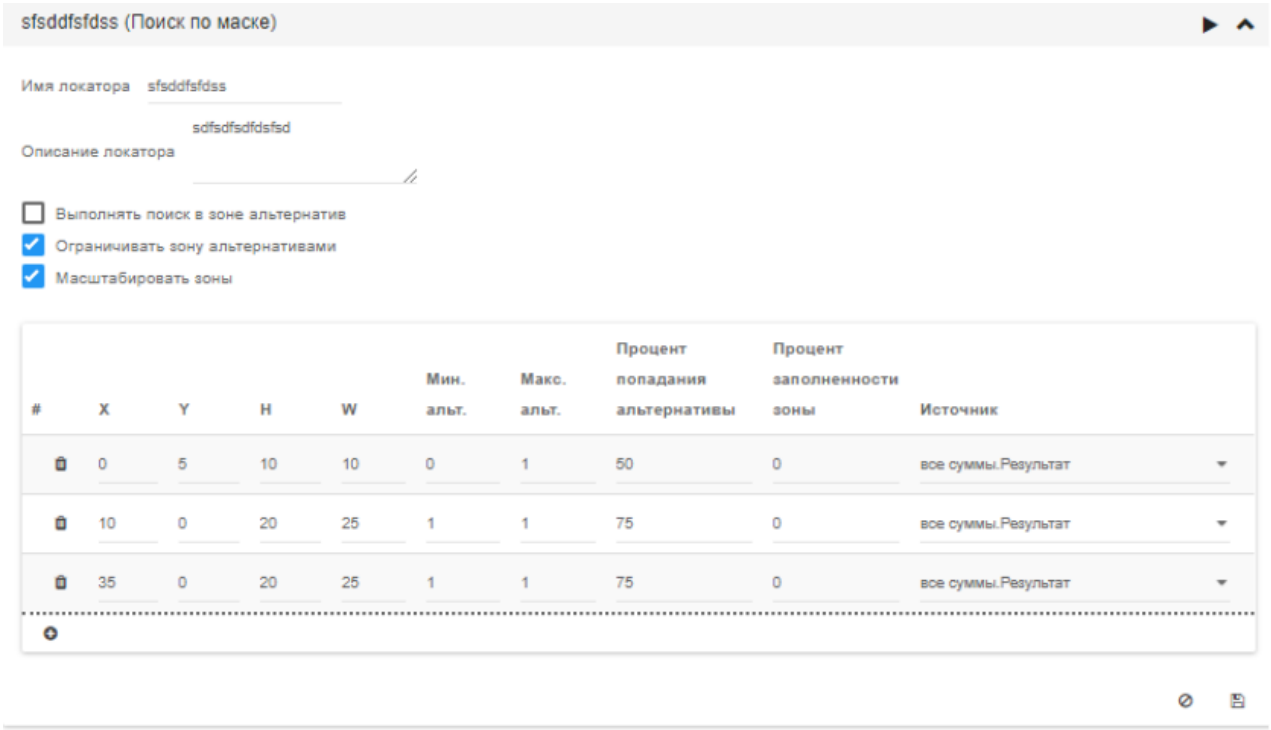
Минимальный процент попадания альтернативы в зону. Указывает минимальный процент области альтернативы, которая накладывается на зону, для того что бы эта альтернатива считалась попавшей в зону. Диапазон значений: 0-100.

Минимальный процент заполненности зоны. Минимальный процент заполнения зоны объединенными альтернативами, попавшими в зону, для того чтобы эта зона была добавлена в качестве ячейки в таблицу. Диапазон значений: 0-100.

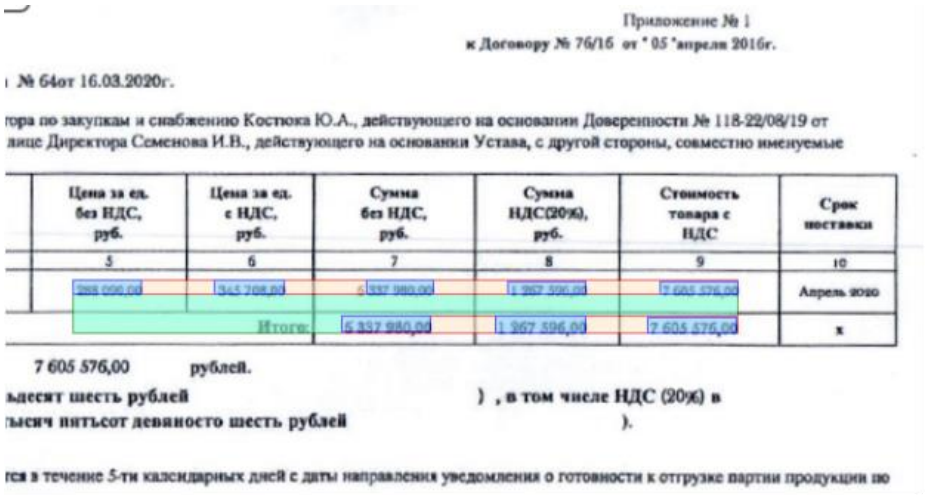
Подполя инструмента извлечения:

Таблица. Таблица, составленная из альтернатив наследуемого подполя, где строка положительный результат наложения маски.

Пример настройки для поиска трех сумм рядом в одном ряду (для всех зон используется одно подполе):



Результат:


Приложение № 1
к Договору № 76/16 от " 05 "апреля 2016г.

№ 64от 16.03.2020г.

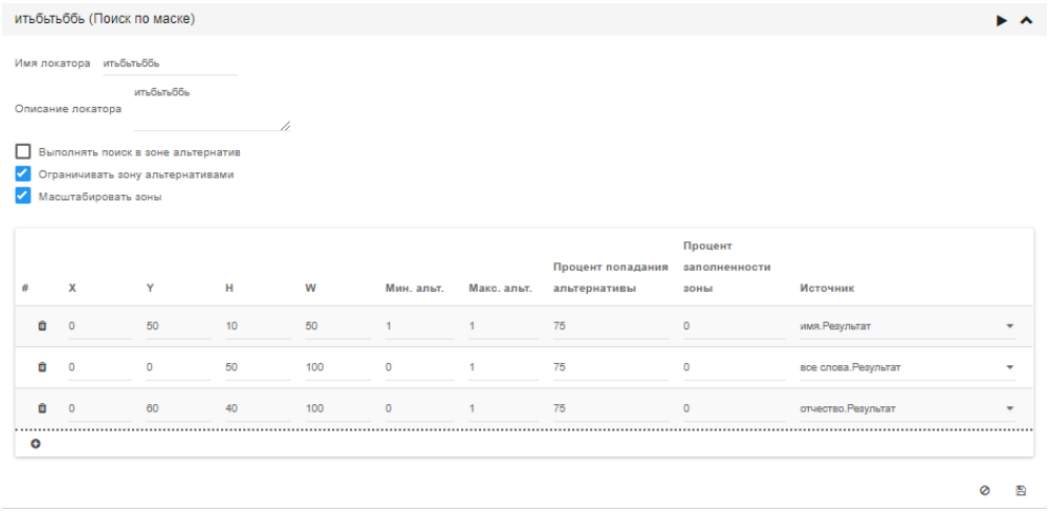
гора по закупкам и снабжению Костюка Ю.А., действующего на основании Доверенности № 118-22/08/19 от
лице Директора Семенова И.В., действующего на основании Устава, с другой стороны, совместно именуемые

Цена за ед. без НДС, руб.	Цена за ед. с НДС, руб.	Сумма без НДС, руб.	Сумма НДС(20%), руб.	Стоимость товара с НДС	Срок поставки
5	6	7	8	9	10
288 000,00	345 708,00	5 337 980,00	1 967 596,00	7 605 576,00	Апрель 2020
	Итого:	5 337 980,00	1 967 596,00	7 605 576,00	x

7 605 576,00 рублей.
ьдесят шесть рублей) , в том числе НДС (20%) в
тысяч пятьсот девяносто шесть рублей).

гся в течение 5-ти календарных дней с даты направления уведомления о готовности к отгрузке партии продукции по

Пример поиска ФИО друг под другом (с разными подполями в зоне):



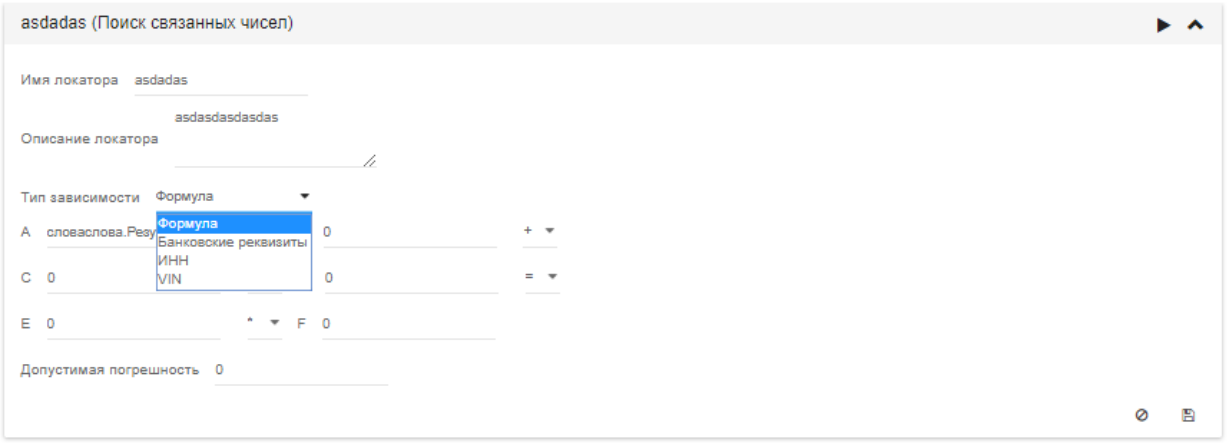
Результат:

31. Поиск связанных чисел ( )

Локатор находит порядок чисел, подходящих указанному условию.

Возможные варианты типа зависимости: Формула, Банковские реквизиты, ИНН, VIN.

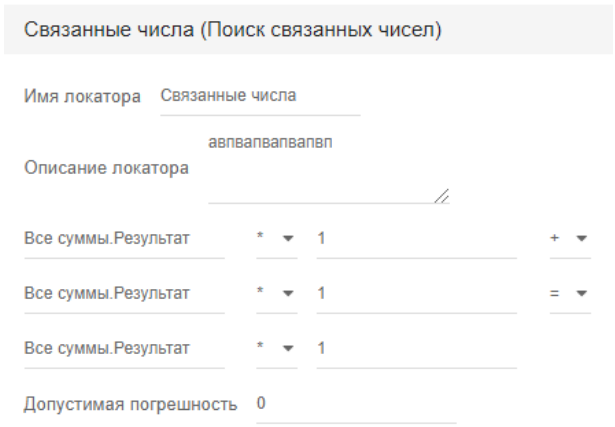
Используются алгоритмы, позволяющие найти данные, для поиска групп связанных банковских реквизитов: БИК, расчётный и корреспондентский счета, а также для поиска ИНН и VIN. Используя данные, основанные на контрольных суммах, можно, например, проверять корректность найденных другими способами данных или выполнять их поиск.



(Рис. 101 Локатор «Поиск связанных чисел»)

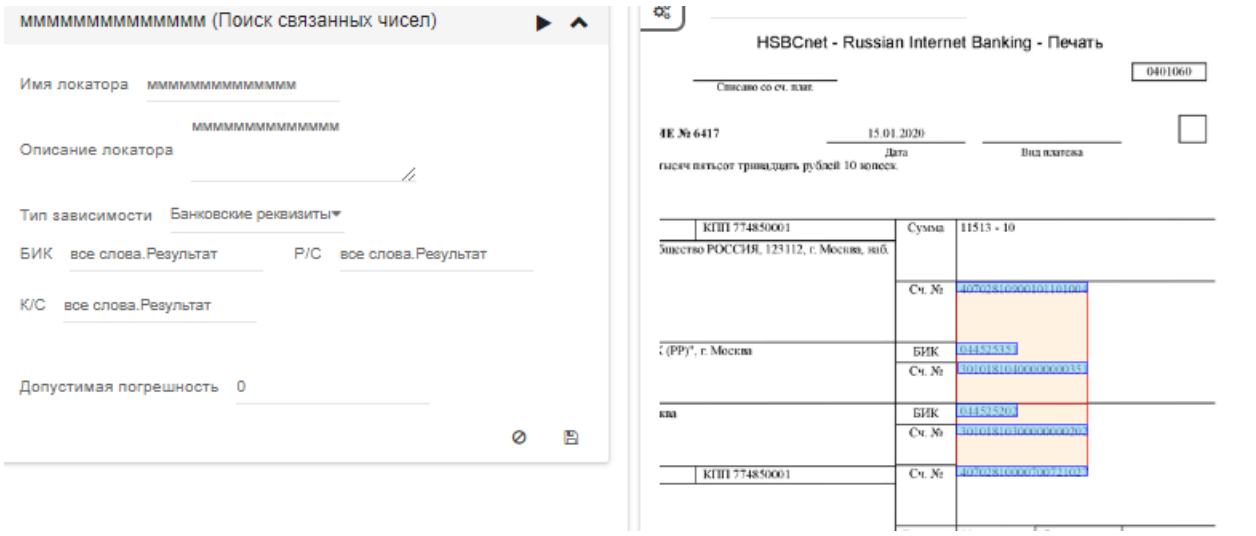
При выборе типа Формула в меню локатора необходимо выбрать результат локатора нахождения чисел и при необходимости коэффициент. Выбрать выполняемую математическую операцию (+ или -) над двумя результатами локаторов и выбрать результат локатора, который будет итогом проведенных операций над числами.

Пример использования типа Формула: На странице находятся все суммы (или все числа) – локатор «Все суммы». В локаторе поиска связанных чисел мы можем найти последовательность цифр: сумма без налога – сумма налога – итоговая сумма. Для этого в локаторе связанных чисел необходимо выбрать локатор «Все суммы» во всех полях, множитель 1, операция сложения:

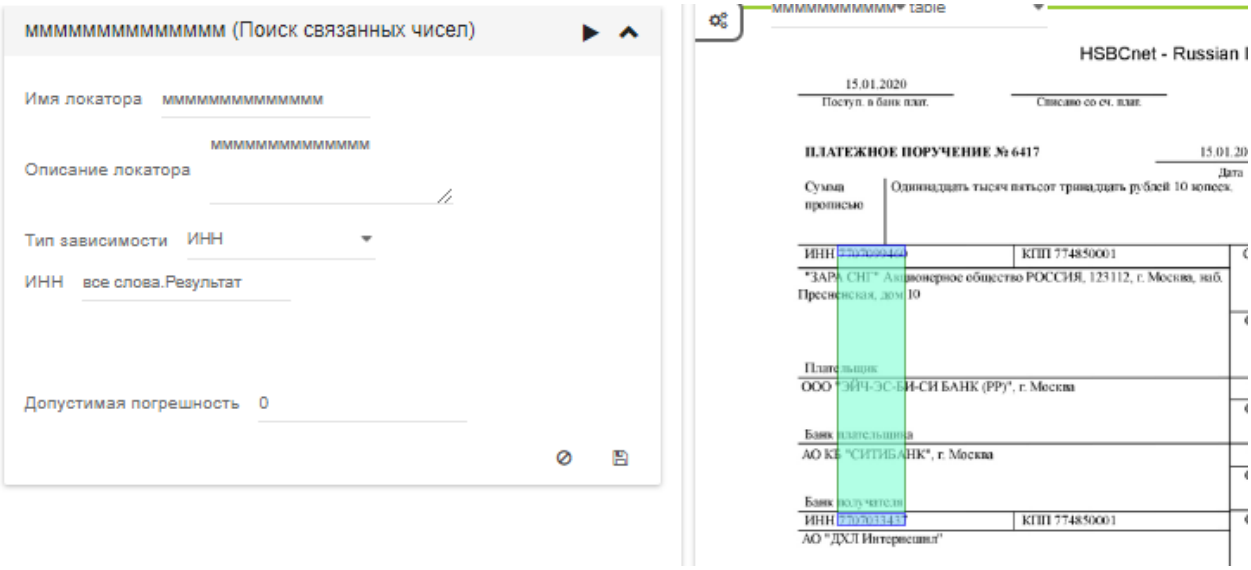


Тогда найдутся все последовательности 3 чисел в которых при сложении двух получается третье.

Пример использования типа Банковские реквизиты: поиск троек БИК, РС, КС:



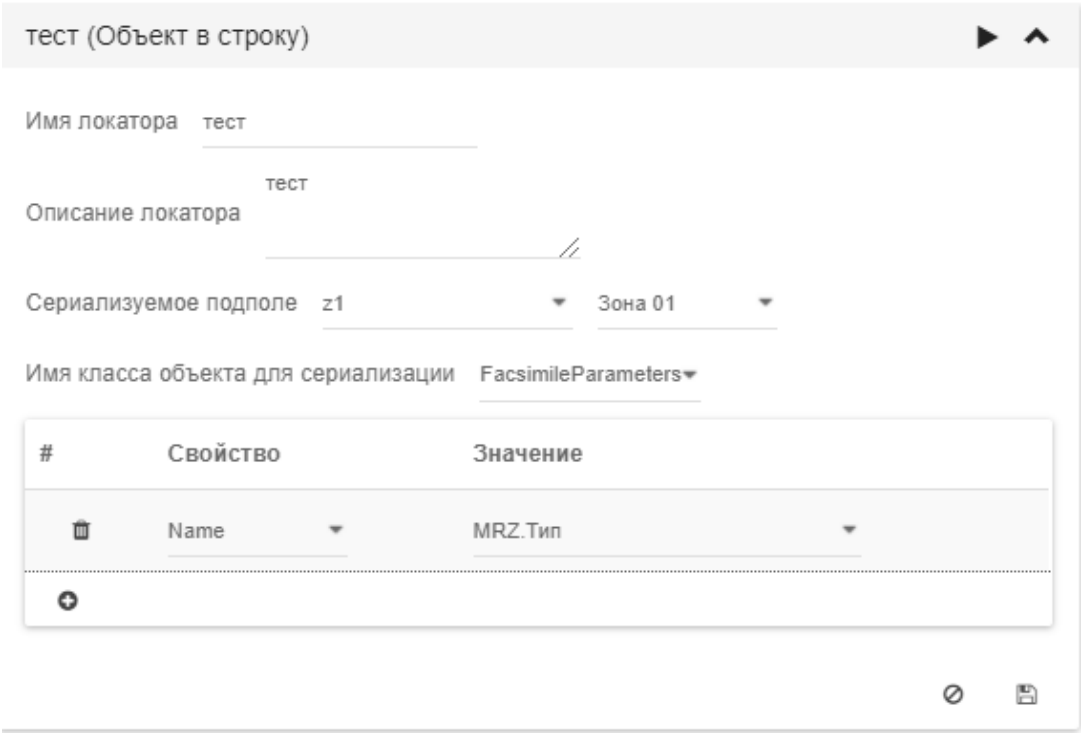
Поиск корректно найденных ИНН:



32. Объект в строку

Локатор предназначен для подстановки в ранее найденную область какого либо изображения из базы. Так же создается XML –код с характеристиками наносимого изображения. При правильно настроенном экспорте XML с параметрами изображения будет выгружаться будет выгружаться.

Чаще всего используется для нанесения факсимиле на экспортируемое изображение.



Сериализуемое подполе - альтернатива найденного заранее локатора в координаты которого будет вставлено изображение. Обязательно выбирается альтернатива область, либо область поиска.

Имя класса объекта для сериализации - класс в зависимости от которого будут выдаваться свойства наносимого изображения.

Свойство – свойства изменяются в зависимости от выбранного класса. Для класса FacsimileParametrs это свойства:

***Name** – имя изображения в базе*

***X,Y** – координаты изображения*

***RotateAngele** – угол наклона изображения*

***Scale** – масштаб изображения*

***Color** – цвет изображения*

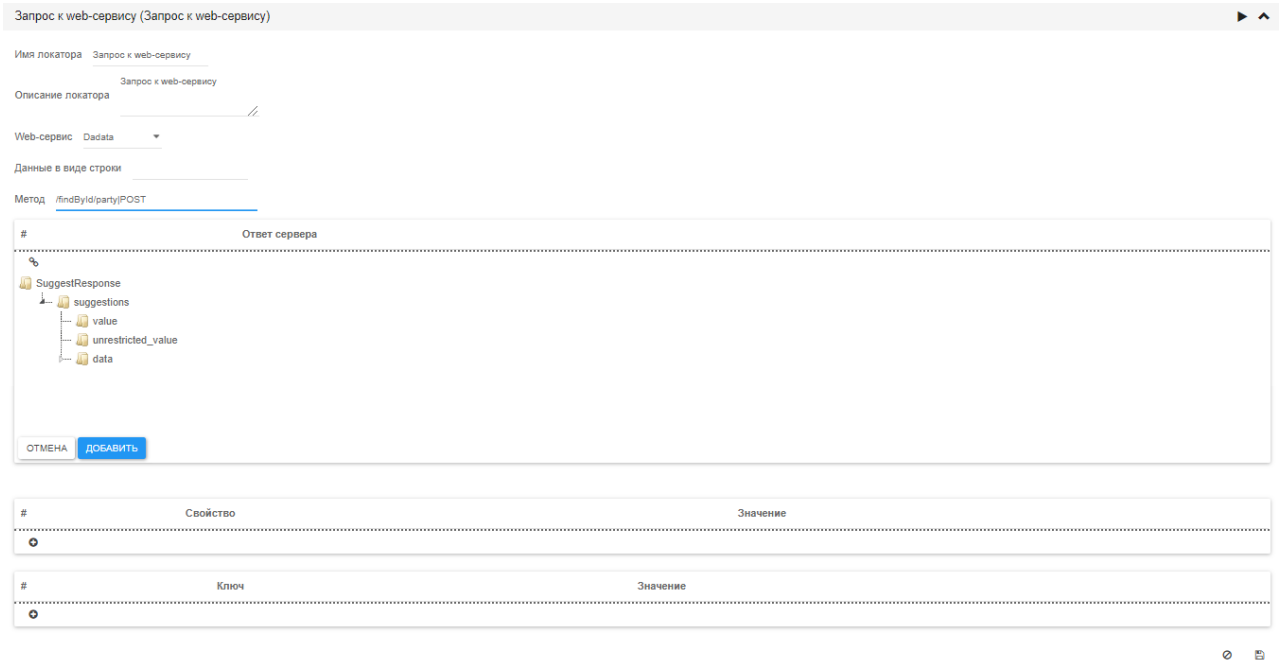
***SpoilPower** – степень сжатия изображения при нанесении.*

Значение – значение выбранного свойства. Может являться альтернативой уже найденного локатора.

33. Запрос к web-сервису

Локатор предназначен для сверки найденных данных на изображении с данными хранящимися на стороннем сервере и по сверенным данным выдача имеющихся в базе.

Например: Локатором на документе находится ИНН. С помощью локатора запрос к web-сервису можно подключиться к сторонней базе, в которой по найденному ИНН можно будет найти Название организации, Адрес организации, КПП и т д. Передавать данные нужно в формате json. Для преобразования, найденного ИНН в формат json можно использовать локатор Объект в строку.



Имя локатора. Указывается имя нового локатора.

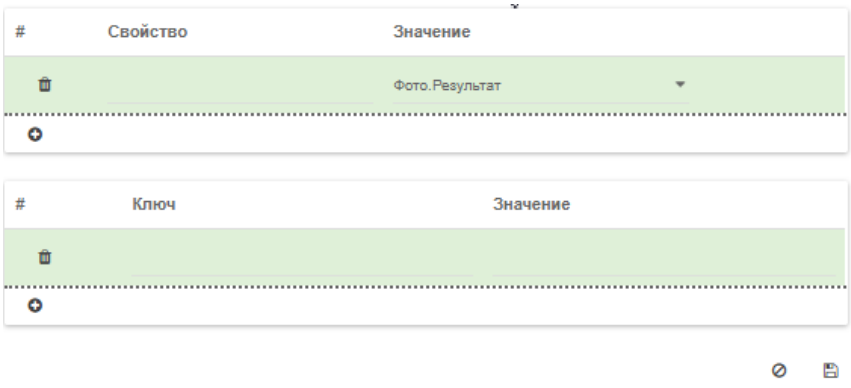
Web-сервис. Из выпадающего списка выбирается сервис, заранее добавленный в веб сервисах. К этому сервису будет обращаться локатор для сравнения данных.

Данные в виде строки. Указывается локатор альтернативы которого будут сравниваться с данными из сервиса. Поле не обязательное. Если поле не заполнено, то в результат будет передаваться весь список по указанным данным из сторонней базы.

Метод. Метод, с которым локатор будет обращаться к сервису. Набор методов должен быть заведомо известен. Набор методов зависит от подключенного сервиса.

Ответ сервера. Окно выбора передаваемых данных. Структура древовидная. Для передачи значения из базы выбирать необходимо value, либо unrestricted_value. Добавленное значение будет отображаться в виде строки. Каждая строка в окне выбора в результате составляет столбец значений.

Для поиска в o_data доступны настройки свойств и ключей.

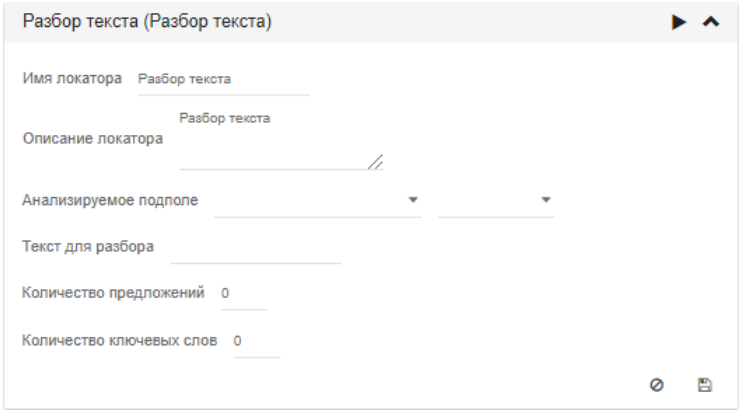


Подполя инструмента извлечения:

Таблица. Таблица, составленная из альтернатив из сторонней базы. Каждый столбец соответствует строке в окне «Ответ сервера».

34. Разбор текста.

Локатор предназначен для выделения ключевых фраз и предложений из общего текста. Выбирается локатор, альтернатива которого будет обработана с помощью нейросети. В результате будет выданы фразы с ключевыми словами и таблица ключевых слов, встречающихся в тексте.



Имя локатора. Имя нового локатора.

Анализируемое поле. Локатор результаты которого будут обработаны настроенной нейросетью для выделения ключевых слов и фраз.

Текст для разбора. Текст для проверки работы инструмента выделения смыла по настроенной нейросети.

Количество предложений. Количество предложений (фраз), которые попадут в результат и будут отражать смысл текста. По умолчанию 2.

Количество ключевых слов. Количество ключевых слов, по которым будет проводится анализ и обработка текста. По умолчанию 10.

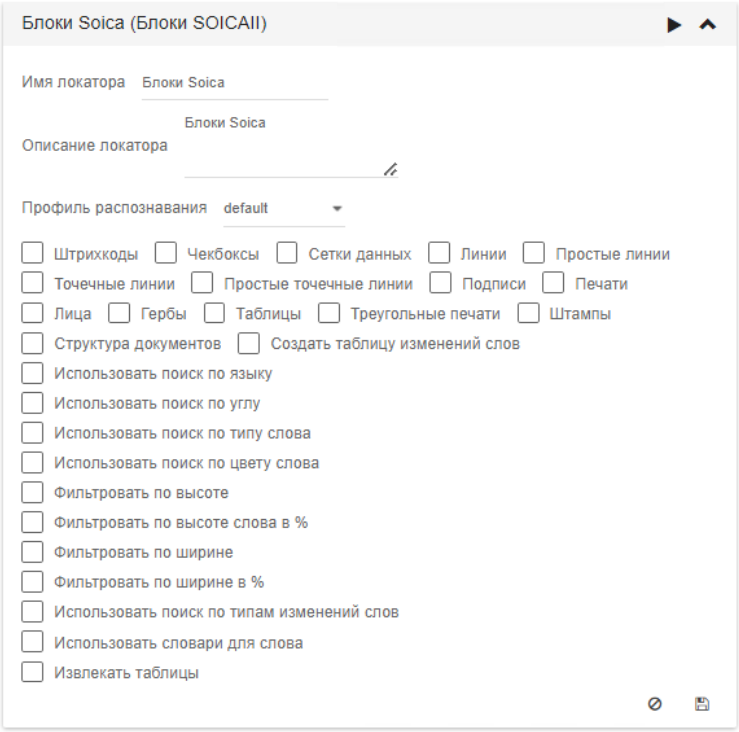
Подполя инструмента извлечения:

Результат. Содержит альтернативы с текстом, фразы, отражающие смысл обрабатываемого текста.

Таблица. Таблица, составленная из найденных ключевых слов и частоты их в тексте.

35. Блоки SOICAII

Локатор предназначен для работы с элементной базой, найденной движком SOICAII.



(Рис.102 Настройка локатора Блоки SOICAII)

В настройках необходимо задать имя, указать профиль и выбрать нужные элементы.

Важно! *Профиль должен быть с движком SOICAII, в котором выбраны нужные опции.*

Подполя инструмента извлечения:

При выборе элемента поиска внутри локатора необходимо чтобы этот же элемент был выбран в указанном профиле распознавании.

Штрих код – найденные штрих коды профилем с движком SOICAII с выбранной опцией «Распознавать штрихкоды».

Чекбоксы – найденные чекбоксы профилем с движком SOICAII с выбранной опцией «Распознавать чекбоксы».

Сетки данных – найденные сетки данных профилем с движком SOICAII с выбранной опцией «Распознавать сетки данных».

Линии, Простые линии – найденные линии профилем с движком SOICAII с выбранной опцией «Распознавать линии».

Точечные линии – найденные точечные линии профилем с движком SOICAII с выбранной опцией «Распознавать точечные линии».

Простые точечные линии – найденные простые точечные линии профилем с движком SOICAII с выбранной опцией «Распознавать точечные линии».

Подписи – найденные подписи профилем с движком SOICAII с выбранной опцией «Распознавать подписи».

Печати – найденные печати профилем с движком SOICAII с выбранной опцией «Распознавать печати».

Лица – найденные лица профилем с движком SOICAII с выбранной опцией «Распознавать лица».

Гербы - найденные гербовые орлы профилем с движком SOICAII с выбранной опцией «Распознавать гербы».

Таблицы - найденные таблицы профилем с движком SOICAII с выбранной опцией «Распознавать таблицы». Выводятся два вида таблиц: таблицы с объединенными ячейками и «классического» вида, где в каждой строке одинаковое количество ячеек.

Треугольные печати – найденные печати профилем с движком SOICAII с выбранной опцией «Искать треугольную печать в центре».

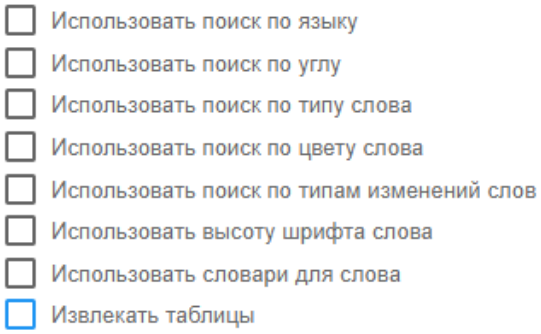
Создать таблицу изменений слов. Позволяет создать таблицу изменений слов при сравнении документов с помощью профиля распознавания с движком DocxComparer.

Структура документов – подполе «hierarchy» содержащее таблицу с 4мя столбцами:

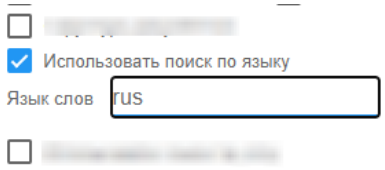
1. Номер параграфа.
2. Ключевые слова параграфа.
3. Сериализованный в формате xml объект параграфа.
4. Текст.

При дальнейшем использовании в локаторах необходимо указывать в локаторе номер столбца, начиная отсчет с «0».

Следующий ряд настроек - это параметры **работы с OCR.** Результат будет записан в альтернативу “word”. При выборе одной из функций будет происходить фильтрация всего полученного OCR.

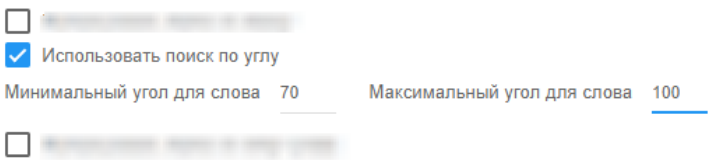


Использовать поиск по языку – при применении этой функции необходимо указать язык выборки. В результате локатора останутся только слова выбранного языка.



Важно! *В настройках применяемого профиля распознавания обязательно должен быть выбран язык, который указывается в локаторе.*

Использовать поиск по углу – при выборе этой функции в альтернативу записываются слова, подходящие под указанный угол наклона. Угол указывается в градусах от 0 до 360.



Использовать поиск по типу слова - при выборе этой функции в альтернативу записываются слова, подходящие выбранному типу. Возможные типы:

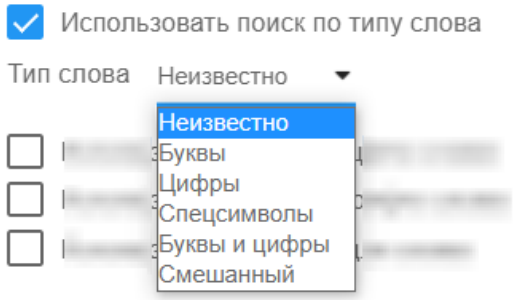
Буквы –только буквы не учитывая регистр;

Цифры – слова, состоящие только из цифр;

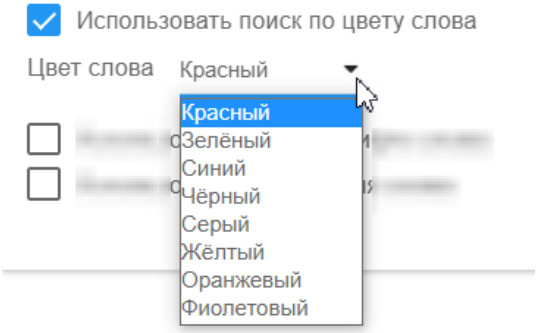
Спец. символы – выборка отдельно стоящих или группы спец. символов.

Буквы и цифры – слова содержащие и буквы, и цифры. Без спец. символов.

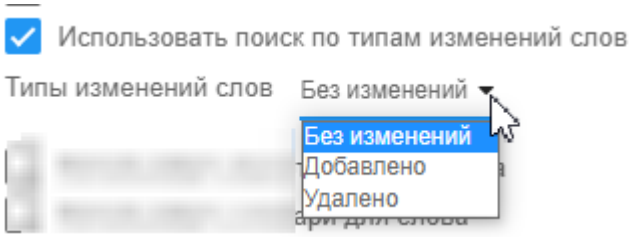
Смешанные – слова с любыми символами.



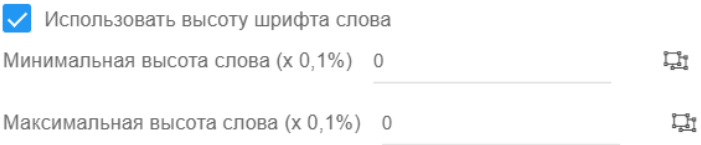
Использовать поиск по цвету слов – при выборе этой функции в альтернативу локатора заносятся только слова выбранного цвета.



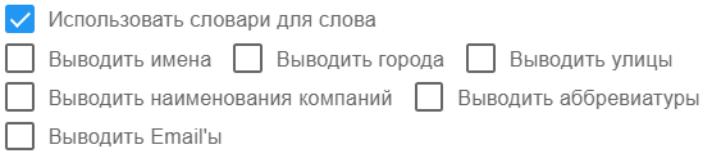
Использовать поиск по типам изменений слов. Данная опция используется при сравнении документов профилем распознавания на базе движка DocxComparer.



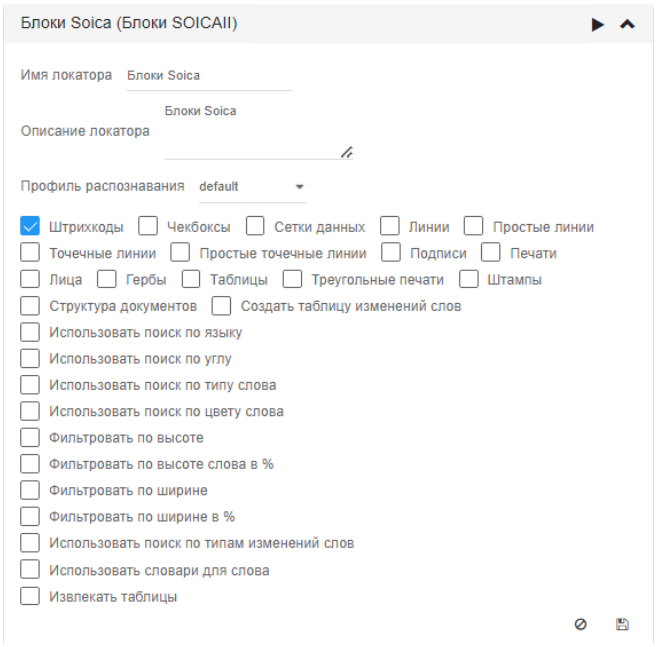
Использовать высоту шрифта слова – при выборе функции в альтернативу попадают только слова, подходящие под указанный размер. Можно использовать для поиска самого маленького или самого большого текста на документе.



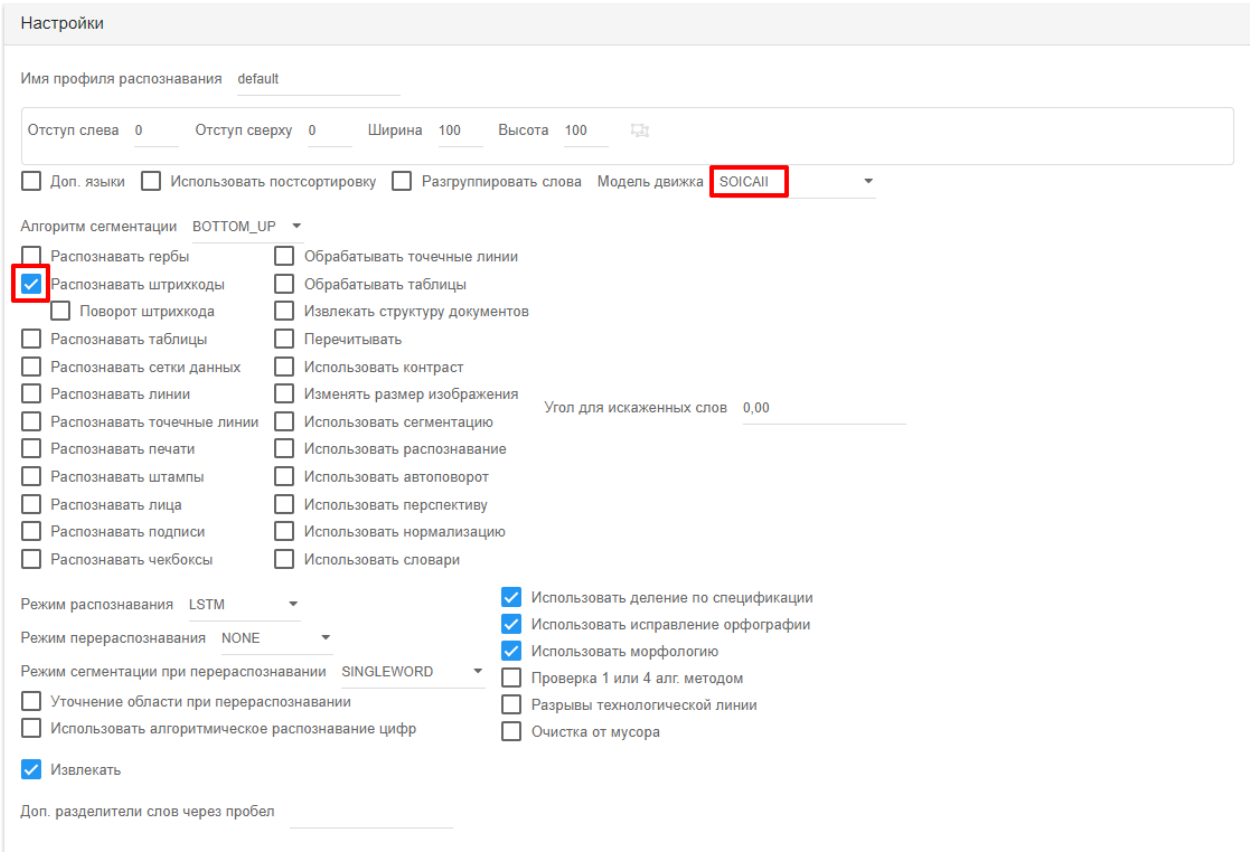
Использовать словари для слова – при выборе этой функции в альтернативу локатора будут попадать только слова, удовлетворяющие выбранным параметрам. Одно слово может удовлетворять нескольким параметрам одновременно. Например, Москва - имя и город.



Для поиска штрих кодов с помощью локатора Блоки SOICA необходимо отметить галочкой «Штрих коды» и убедиться, что в выбранном профиле распознавания на базе движка SOICAII выбрана соответствующая опция.

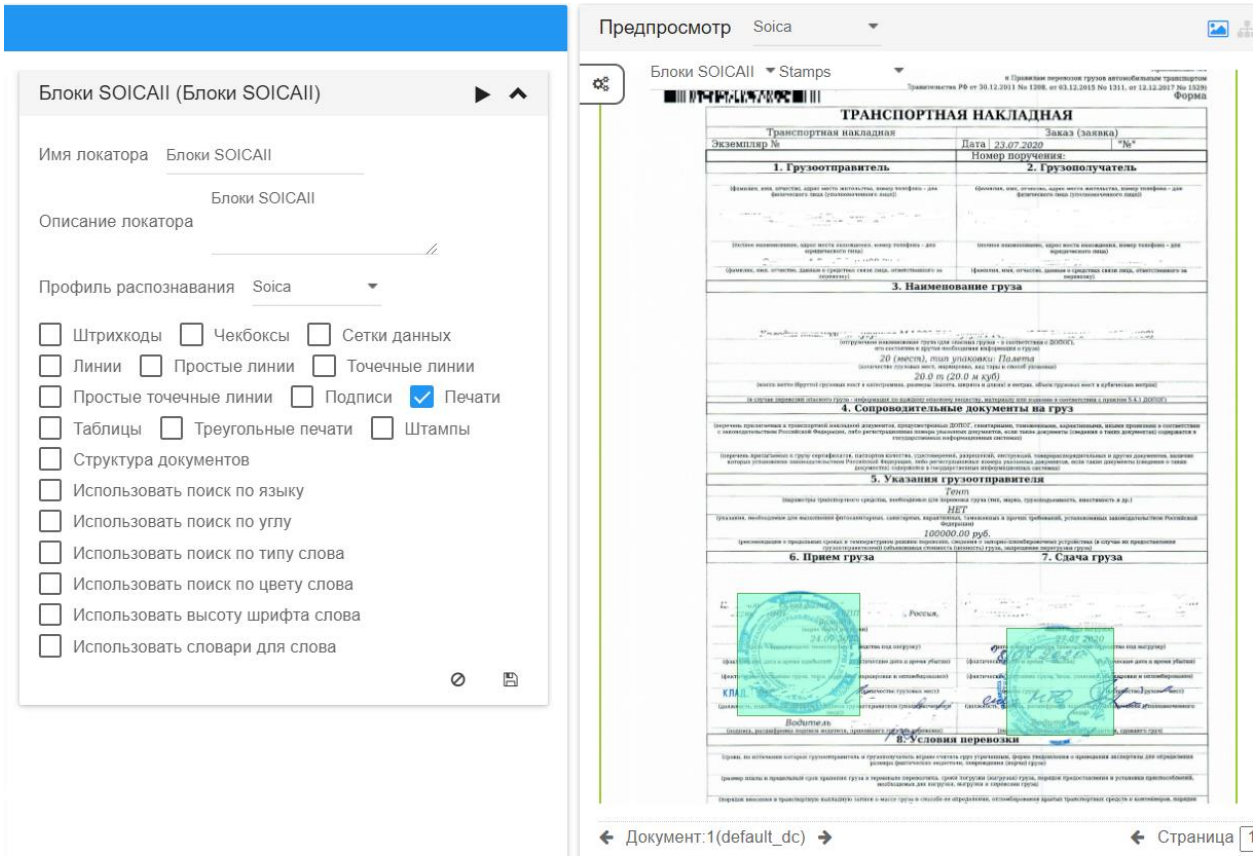


(Рис.102.1 Настройка локатора Блоки SOICAII для поиска Штрих кодов)

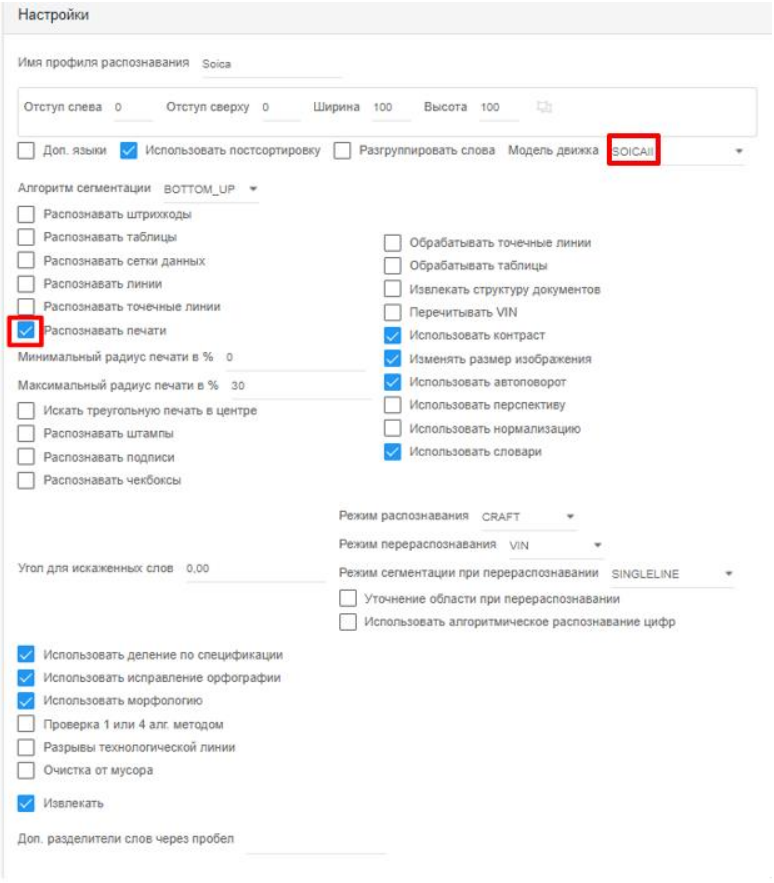


(Рис.102.2 Настройка профиля распознавания с движком SOICAII для поиска штрих кодов)

Для поиска печатей с помощью локатора Блоки SOICA необходимо отметить галочкой «Печати» и убедиться, что в выбранном профиле распознавания на базе движка SOICAII выбрана соответствующая опция.

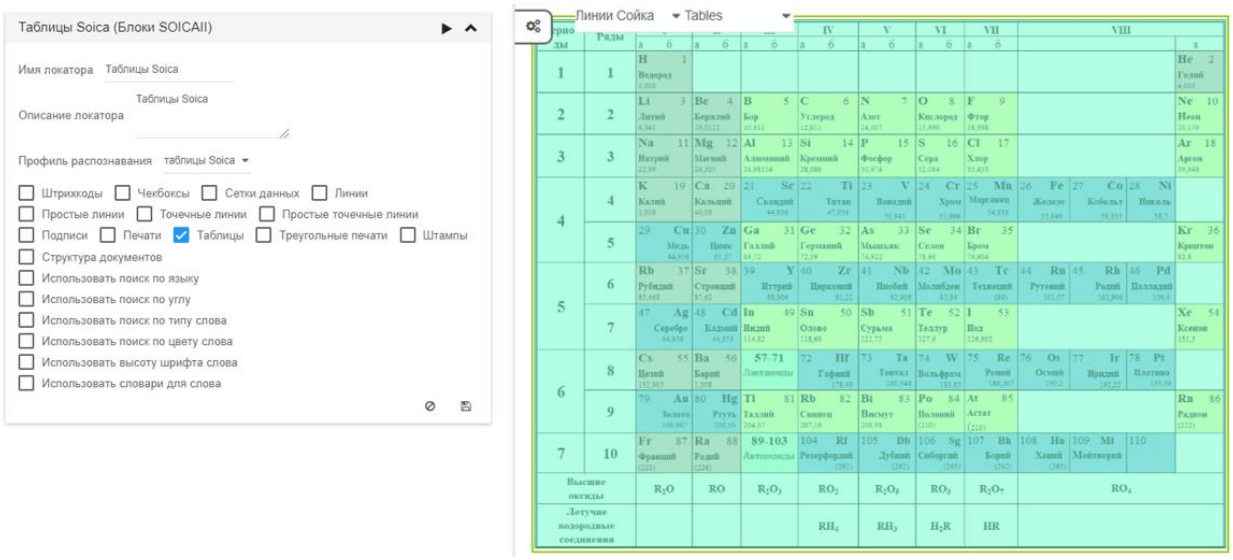


(Рис.102.3 Настройка и результат локатора Блоки SOICAII для поиска печатей)

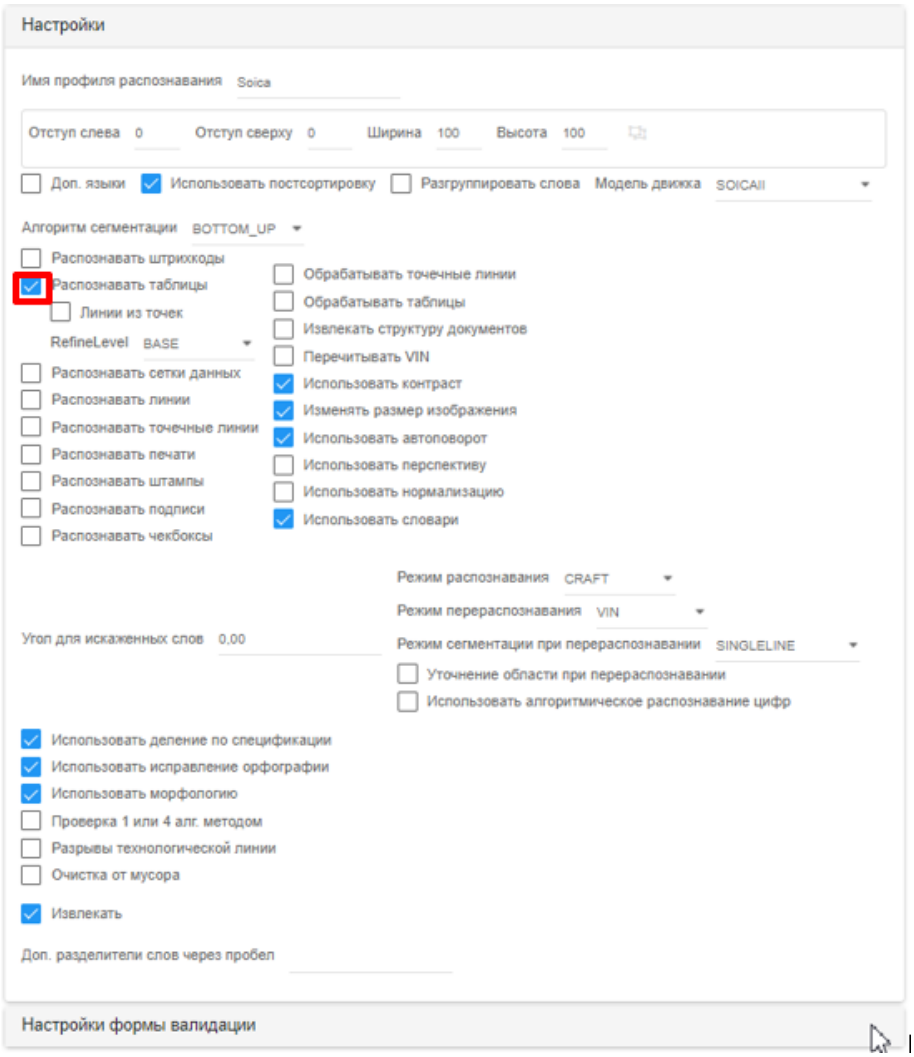


(Рис.102.4 Настройка профиля распознавания с движком SOICAII для поиска печати)

Для поиска таблицы с помощью локатора Блоки SOICA необходимо отметить галочкой «Таблицы» и убедиться, что в выбранном профиле распознавания на базе движка SOICAII выбрана соответствующая опция.

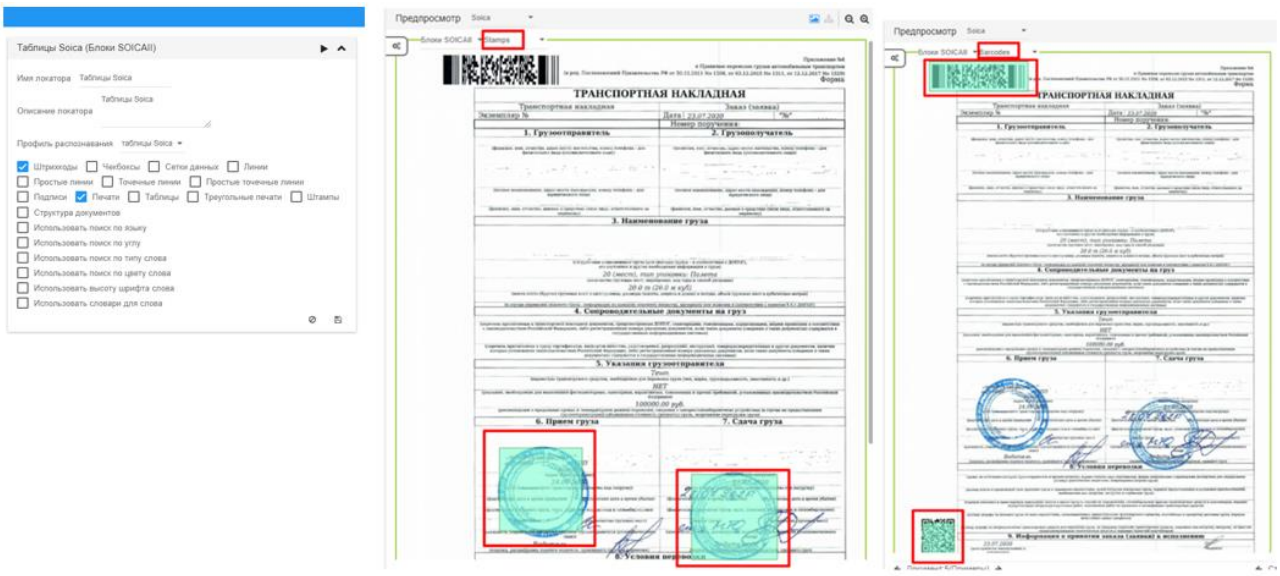


(Рис.102.5 Настройка и результат локатора Блоки SOICAII для поиска таблицы)

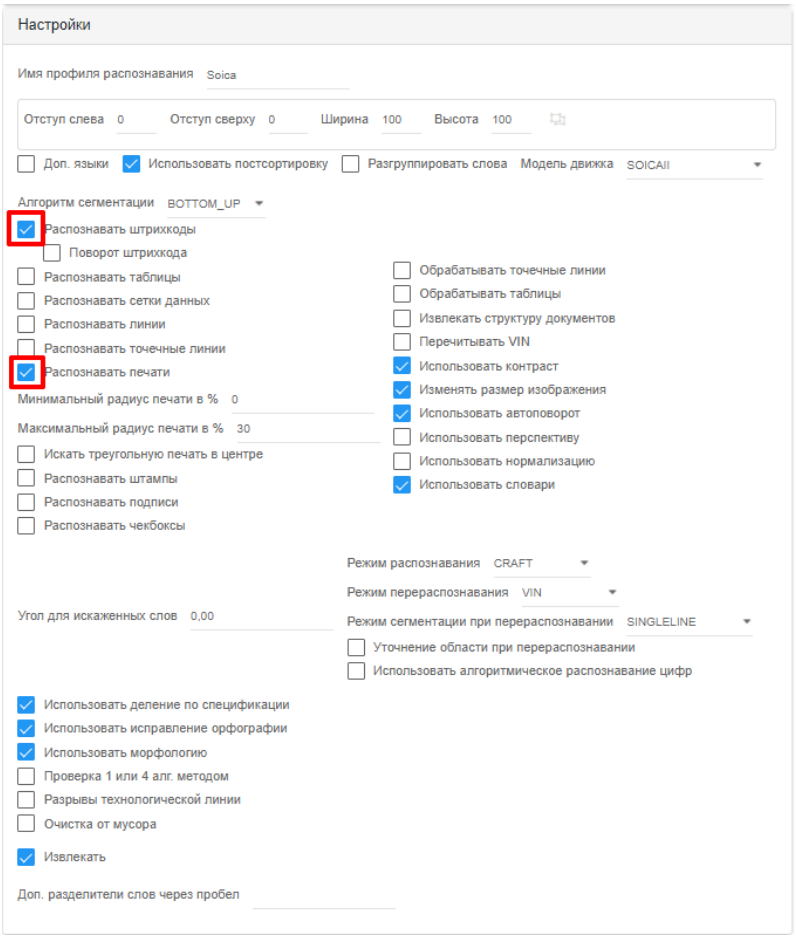


(Рис.102.6 Настройка профиля распознавания с движком SOICAII для поиска таблицы)

Для одновременного поиска нескольких типов данных с помощью локатора Блоки SOICA необходимо отметить галочкой нужные опции и убедиться, что в выбранном профиле распознавания на базе движка SOICAII выбраны соответствующие опции.



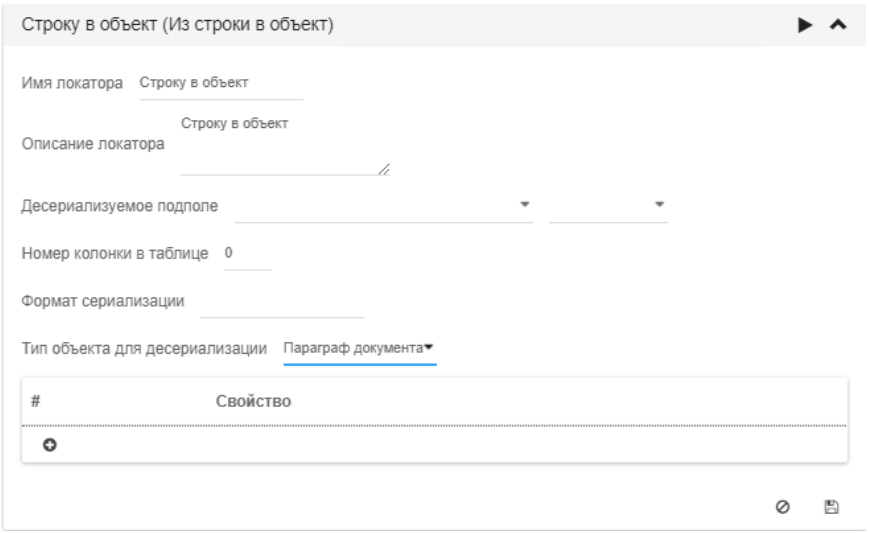
(Рис.102.7 Настройка и результаты локатора Блоки SOICAII для поиска печатей и штрих кодов)



(Рис.102.8 Настройка профиля распознавания с движком SOICAII для поиска штрих кодов и печатей)

36.Строка в объект

Локатор используется для того, чтобы извлечь из структуры документа необходимые данные.



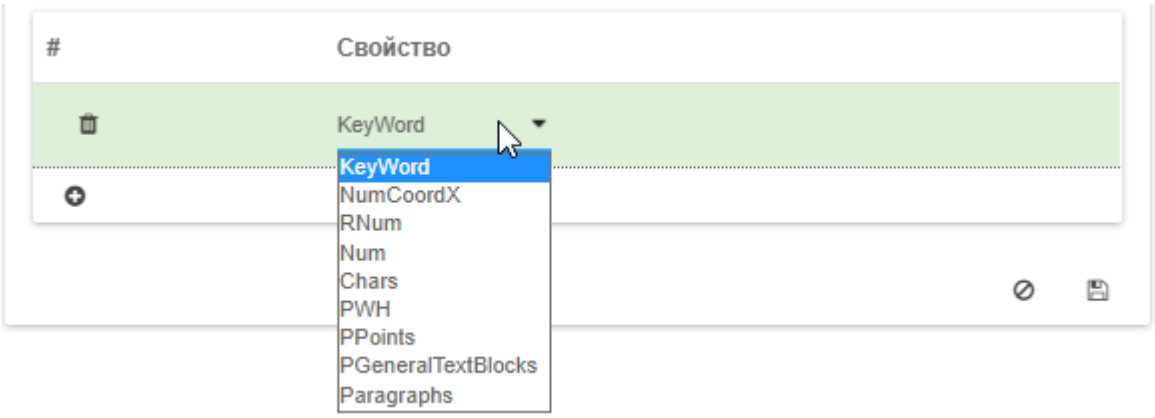
«Десериализуемое подполе» - необходимо выбрать результаты локатора «Блоки SOICA», который извлекает структуру документа.

«Номер колонки в таблице» - номер колонки из результатов локатора «Блоки SOICA», содержащий в себе сериализованный в xml объект параграфа

«Формат сериализации» - выбор формата для сериализации (xml)

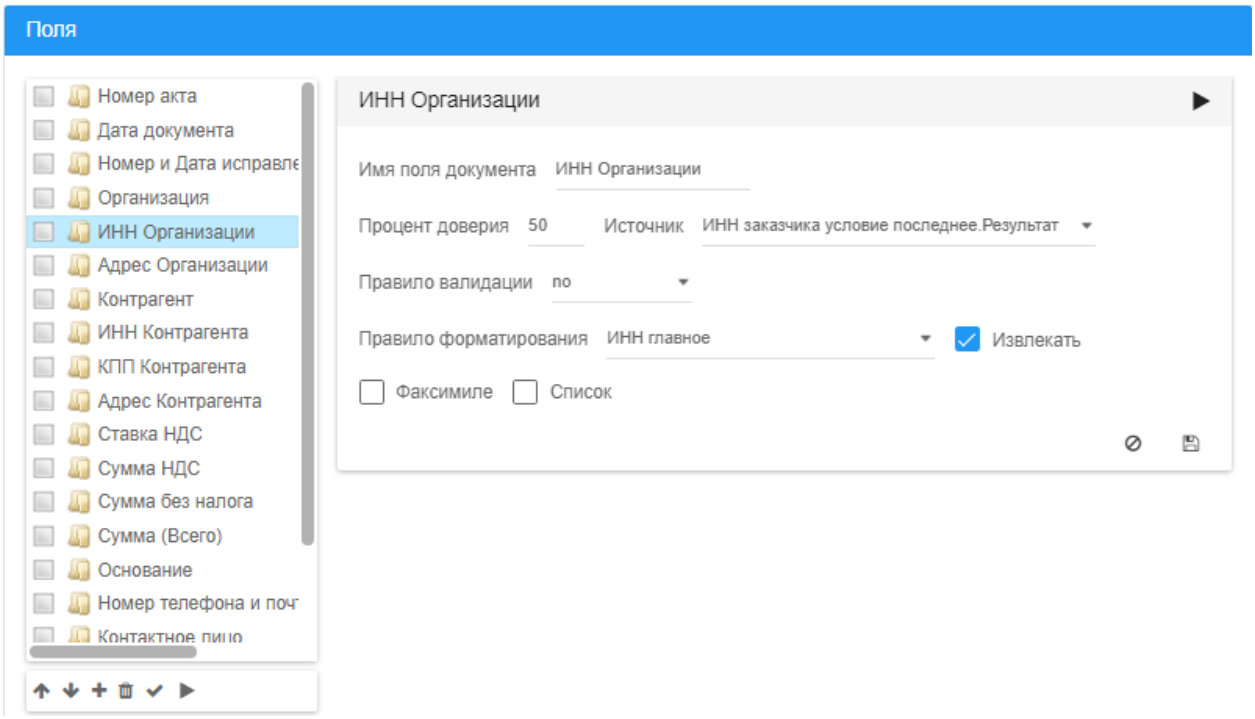
«Тип объекта для десириализации» - параграф документа

«Свойство» - выбор свойств для результирующей таблицы



2.4.2 Поля.

После получения данных из документа при помощи локаторов. Результаты работы локаторов можно присваивать полям, предварительно созданным администратором. Результатом поля может быть результат любого локатора, количество полей не ограничено. Так же к полям можно применять правила форматирования и валидации. В дальнейшем данные из полей будут экспортированы в целевую систему.



(Рис. 103 Интерфейс настройки полей)

Область настройки полей делится на две зоны: Список выбора поля и настройки выбранного поля.

В зоне настройки выбранного поля можно:

Имя поля документа. Изменяет имя, отображаемое в списке полей.

Процент доверия. Указывается степень доверия к найденным данным. Если у найденных данных процент будет ниже указанного, то при валидации данное поле будет подсвечено красным, если выше или равно, то зеленым.

Источник. Результат выбранного подполя локатора.

Привило валидации. Правило валидации применимое к данному полю.

Привило форматирования. Правило форматирования применимое к данному полю.

Извлекать. Если отметка стоит, то поле будет отрабатывать и попадет на валидацию.

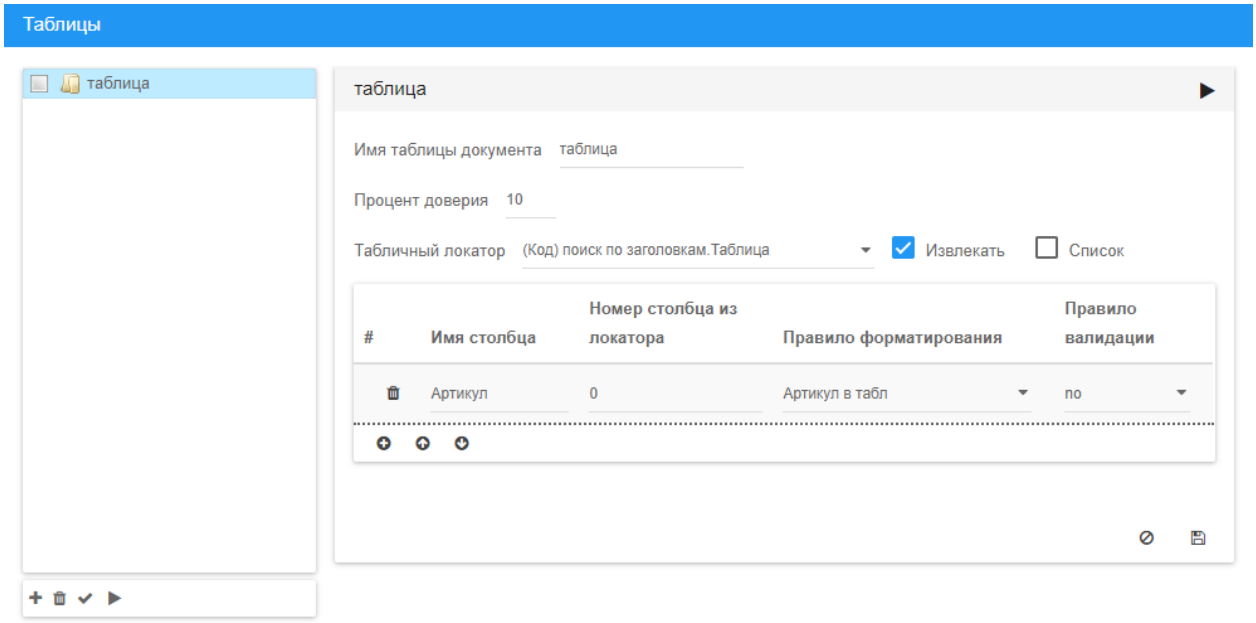
Факсимиле. Обозначает что выделено поле является Факсимиле. Поля с этой отметкой по-другому проходят этап валидации.

Список. Результирующие альтернативы данного поля будут выводиться в виде списка.

Результаты работы полей можно увидеть в предпросмотре в древовидном виде.

2.4.3 Таблицы.

Для вывода результатов работы табличного локатора используется специальная форма «Таблица». Область «Таблицы» делится на две зоны: Список таблиц и настройка выбранной таблицы.



(Рис. 104 Таблицы)

Имя таблицы документа. Имя, которое будут отображаться в списке таблиц.

Процент доверия. Указывается степень доверия к найденным данным. Если у найденных данных процент будет ниже указанного, то при валидации ячейка таблицы будет подсвечена красным, если выше или равно, то зеленым.

Табличный локатор. Выбирается локатор с табличными подполями. В таблице можно выбрать только табличные локаторы, а в полях нельзя.

Извлекать. Если отметка стоит, то таблица будет отрабатывать и попадет на валидацию.

Список. Если данная опция выбрана, то на форме валидации будут выводиться все таблицы, являющиеся результирующими.

Имя столбцы. Имя столбца, которое будет отображаться при валидации и экспорте.

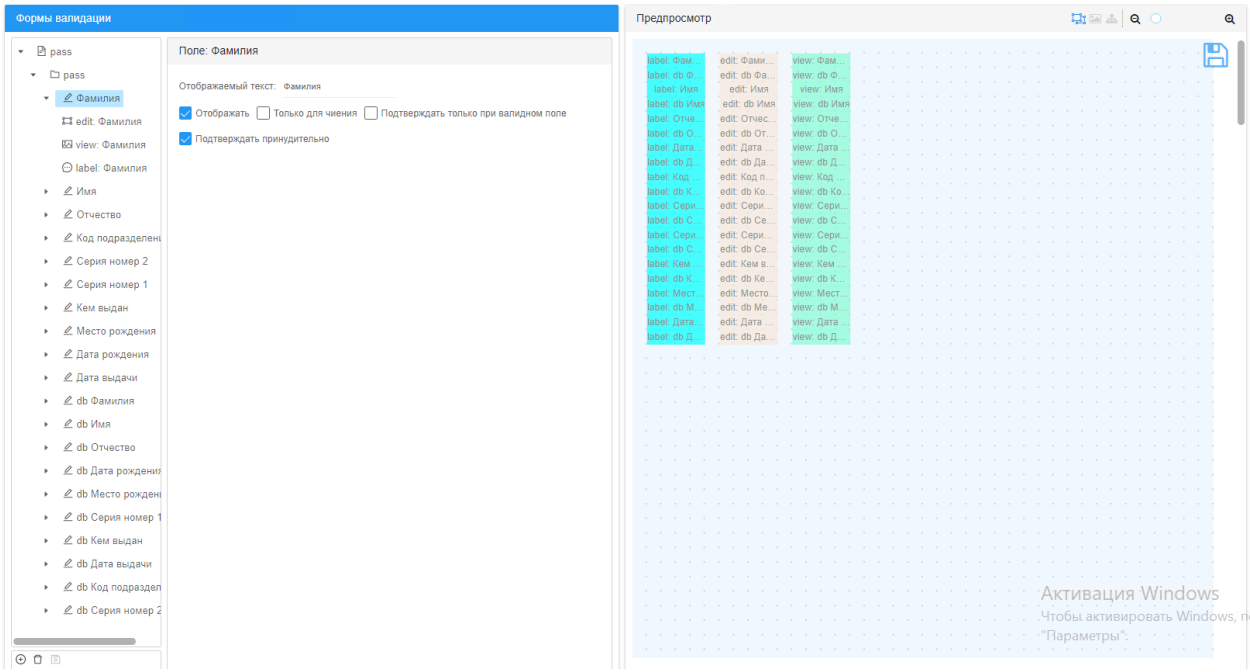
Номер столбца из локатора. Номер столбца из выбранного табличного локатора. Нумерации начинается с 0.

Правило форматирования. Правило форматирования применимое к данному столбцу.

Правило валидации. Правило валидации применимое к данному столбцу.

2.4.4 Форма валидации.

Данная форма используется для настройки визуального отображения полученных результатов на этапе валидации (проверки извлеченных данных оператором). Необходимость и условия перехода на этап валидации настраиваются администратором. Форма валидации делится на две части: Список элементов формы (в виде дерева) и Визуальное отображение формы.

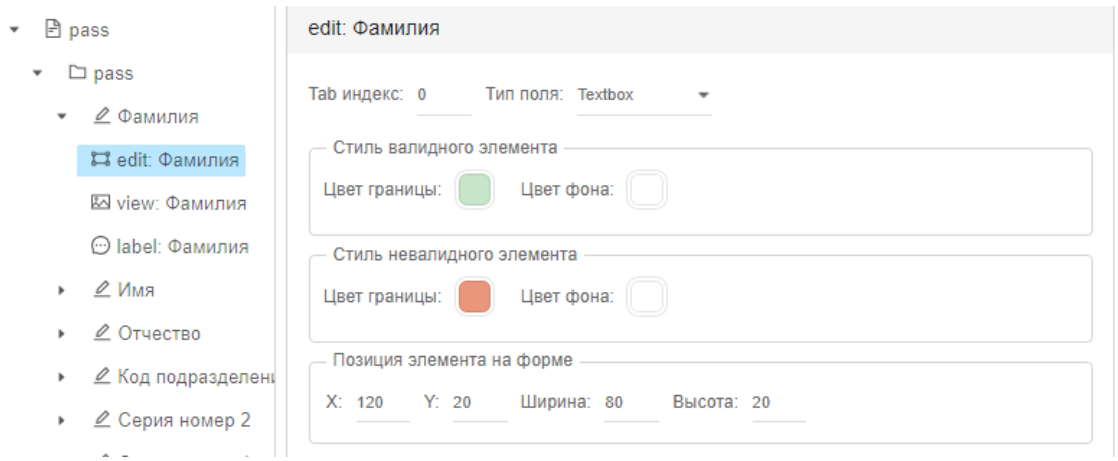


(Рис. 105 Форма валидации)

В списке элементов можно отдельно корректировать все настройки каждого элемента.

На визуальной форме можно перемещать и изменять поля вывода данных.

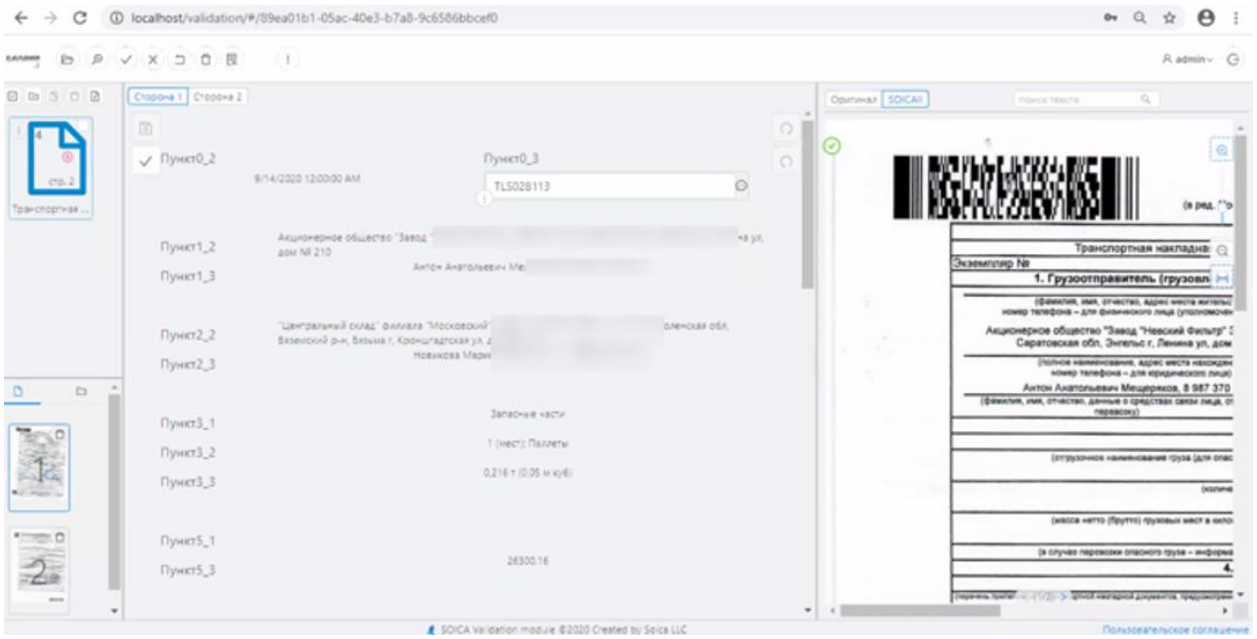
В области древовидного отображения полей можно добавлять вкладки на форму валидации. Перемещать поля на новую вкладку можно «перетаскиванием» нужного поля в дереве объектов на новую вкладку.



На форме Валидации в текстовом блоке «Text box» есть возможность выставить параметр поля «только чтение».

Если данный параметр выставлен, то поле становится не редактируемое на Валидации.

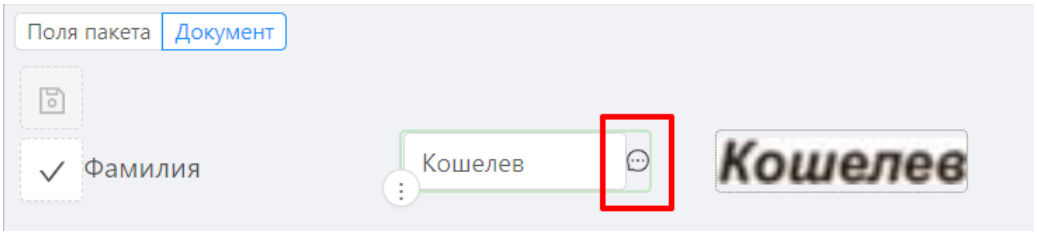
Таким образом в модуле Валидации выглядит поле с включенным параметром «Только чтение».



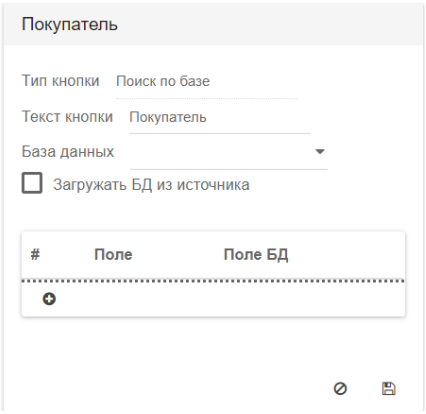
При этом, если поле привязано в настройках кнопки Валидации, то напротив него не появляется кнопка поиска по БД и на форме поиска оно тоже не отображается.

2.4.5 Кнопки валидации.

Кнопки валидации отображаются на форме валидации в правой части полей, к которым они привязаны.



Они предназначены для обращения оператора к некоему источнику данных и для смены пользователя.



При создании кнопки указывается имя кнопки и текст, который будет отображаться на форме валидации.

Указывается база данных. Выбирается внешний источник и з выпадающего списка.

При установке отметки «Загружать БД из источника» при каждом обращении к кнопки данные внешнего источника будут обновляться из фала на сервере.

В полях указывается соответствия отображаемого поля и поля БД.

(Рис. 106 Настройка кнопки валидации)

Если кнопка Валидации привязана к базе, которая в свою очередь является БД PG или MS, при этом стоит на кнопке валидации галочка «загружать БД из источника», тогда на форме Валидации при выполнении поиска по базе используя кнопку с данными условиями будет вытягиваться не вся таблица из источника, а только та, которая удовлетворяет условиям поиска.

2.4.6 Переклассификация.

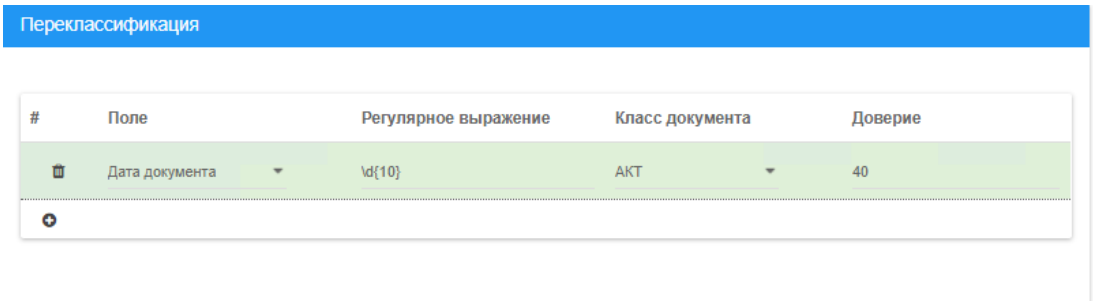
Переклассификация - дополнительный этап классификации который выполняется после извлечения данных.

На основе извлеченных данных и заданных условий в настройках переклассификации, документу может присвоиться другой класс, после чего он будет обрабатываться другой логикой и будет участвовать в других маршрутах экспорта.

В случае нахождения текстовых данных переклассификация может реагировать на соответствие текста заданному регулярному выражению.

- Поле – поле документа по которому будет выполняться переклассификация
- Регулярное выражение – если значение поля соответствует указанному регулярному выражению, то для этого документа указывается кандидат класса документа с указанной степенью доверия.

После проверки всех строчек переклассификации для документа выбирается результирующий класс, с большей суммарной степенью доверия.



(Рис. 107 Настройка переклассификации)

2.4.7 Создание документов из изображения поля.

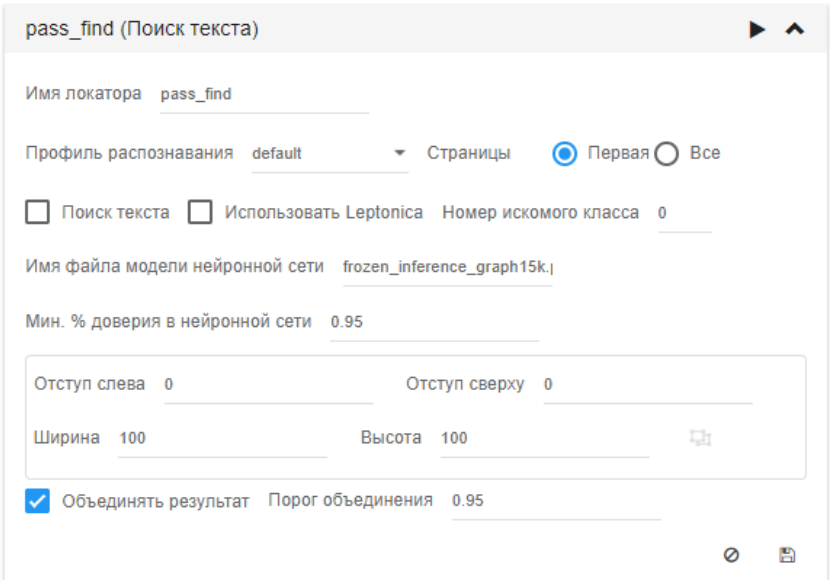
Возможно в процессе распознавания создавать одностраничные документы из части изображения страницы имеющегося документа. Для этого необходимо унаследовать нужную часть изображения страницы документа полю документа.

В результатах полей могут быть как текстовые, так и графические данные.

Для этого можно использовать поля, полученные из локатора настроенного с помощью использования натренированных нейросетей. В локаторах используются модели TensorFlow. Выбор модели происходит на уровне настройки локатора. Указывается поле, из изображения которого будет создаваться документ указанного класса.

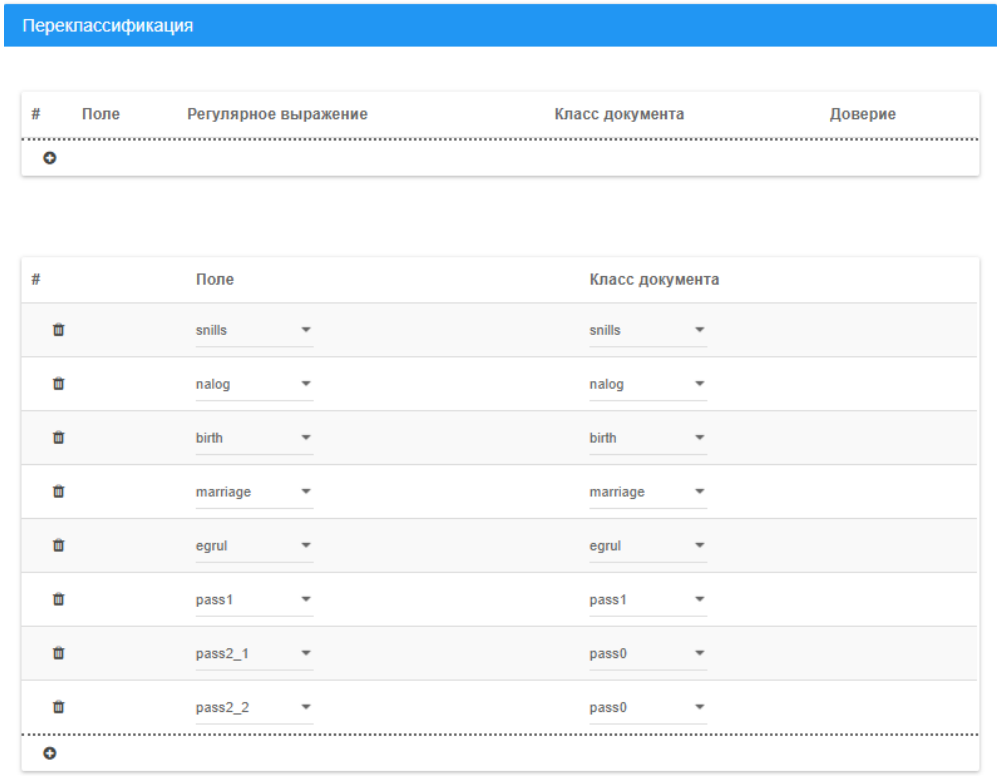
Модель нейросети может быть натренирована под любую задачу. Нахождение и выделение области может быть, как для ценника или штрих-кода, так и для паспорта или другого ДУЛа.

Пример:



Настройка локатора «Поиск на основе нейронной сети». В качестве натренированной модели TensorFlow здесь используется расположенный на сервере файл: «frozen_inference_graph15k». Он ищет класс с системным номером 0 (Номер искомого класса 0).

Если данная область была найдена на документе, то в поле pass1 передаётся значение локатора.



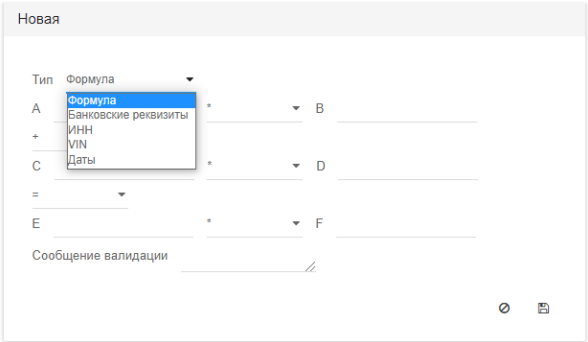
На снимке экрана мы видим, что из поля pass1 в столбце «Поле» если оно заполнено, создается документ класса «pass1», указанного в столбце «Класс документа».

Таким образом мы можем присваивать классы документов отдельным частям исходного изображения. Данная операция происходит после общей предобработки и распознавания. Новый выделенный документ поступает на этап распознавания и к нему применятся все настройки нового класса документа.

2.2.9. Групповая валидация.

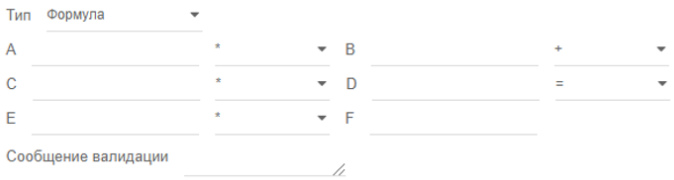
Групповая валидация предназначена для проверки результатов полученных значений.

В правилах групповой валидации сравниваются результаты полей, полученных ранее.

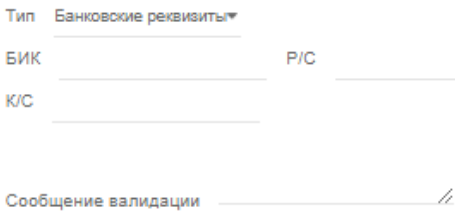


(Рис. 108 Настройка групповой валидации)

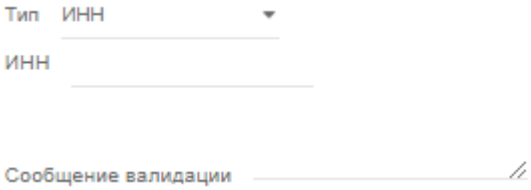
Доступны 5 типов проверки: Формула, Банковские реквизиты, ИНН, VIN, Дата



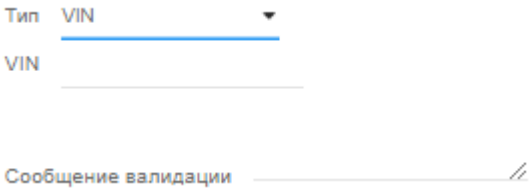
(Рис. 108.1 Типы проверки и настройка типа «Формула»)



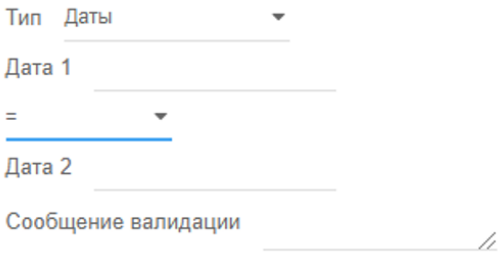
(Рис. 108.2 Настройки типа «Банковские реквизиты»)



(Рис. 108.3 Настройки типа «ИНН»)



(Рис. 108.4 Настройки типа «VIN»)



(Рис. 108.5 Настройки типа «Дата»)

Например, мы можем проверить итоговую сумму с НДС путем прибавления 20% к сумме без НДС, если валидация успешна – результаты верны.

Принцип работы схож с принципом работы локатора «Связанные числа».

Сообщение валидации. Необходимо написать сообщение, которое будет появляться при невалидности.

3. Модуль Валидации.

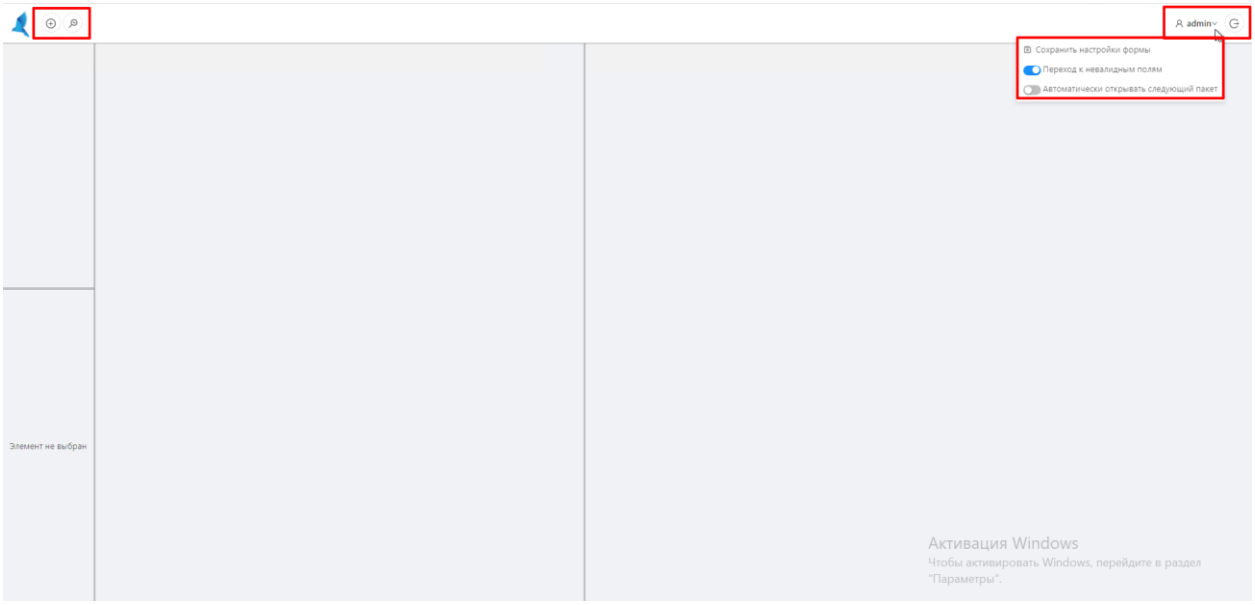
Модуль предназначен для работы оператора валидации. В модуле происходит проверка полученных данных после распознавания. Данные, которые система считает не валидными будут подсвечиваться красным.

3.1. Вход в модуль

Для входа в главное меню модуля пользователю необходимо ввести свой уникальный логин и пароль в соответствующие окна. Далее необходимо нажать кнопку «Войти».

3.2. Главный экран модуля

После того, как пользователь нажал кнопку «Войти» он переходит в главное меню модуля. В левом верхнем углу расположены 2 кнопки с различным функционалом, рассмотрим каждую из них по отдельности.

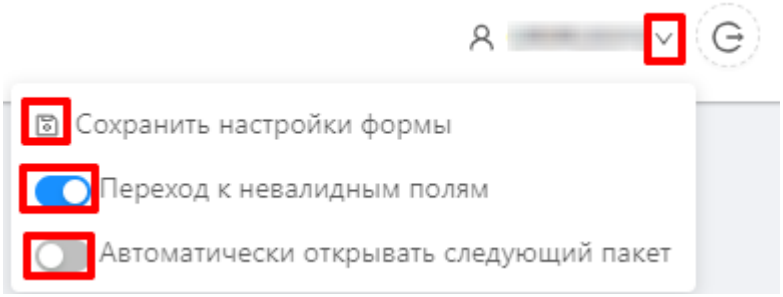


- Создать новый.

- Монитор статуса пакетов.

В правом верхнем углу отображается логин пользователя и кнопка, отвечающая за выход в окно ввода логина и пароля .

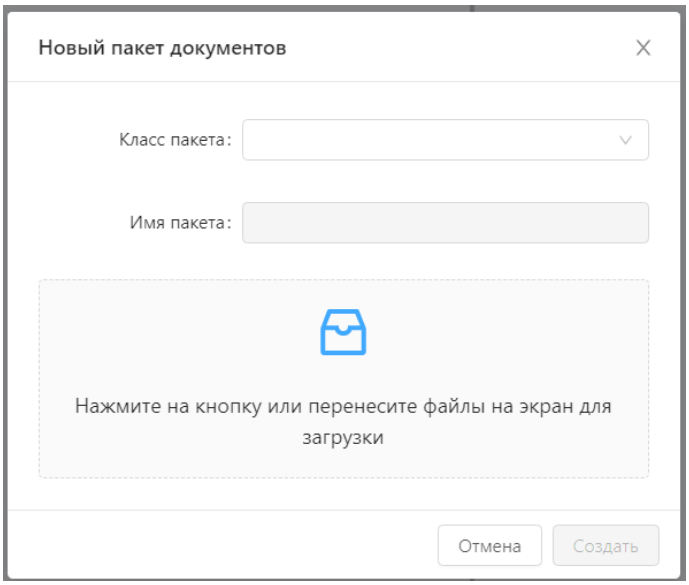
Справа от логина пользователя вызов кнопки сохранения настройки формы для данного пользователя и кнопка выбора варианта перехода между не валидными полями. Если опция «Переход к не валидным полям» включена, то при подтверждении не валидного поля системой автоматически будет совершен переход к следующему не валидному полю. Если опция «Автоматически открывать следующий пакет» включена, после подтверждения открытого пакета на форме валидации, автоматически откроется следующий по списку пакет.



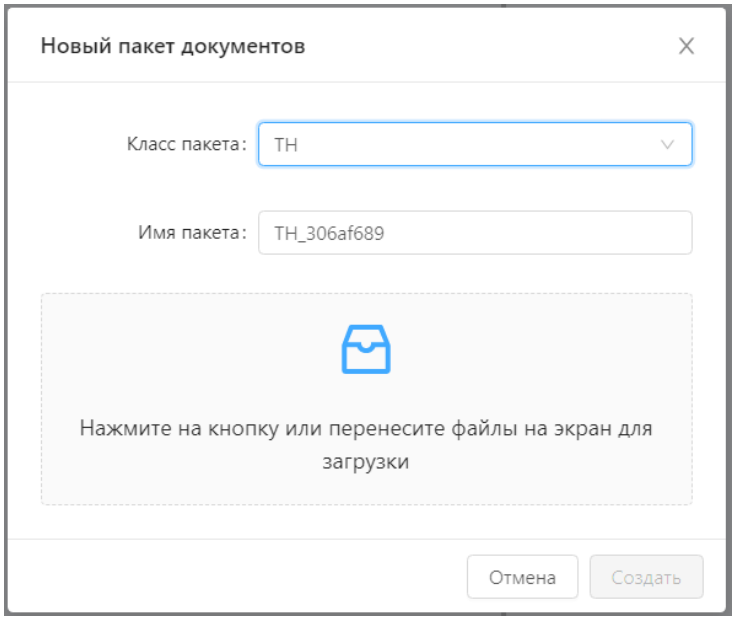
3.3. Функционал кнопки «Создать новый»

«Создать новый» →

Позволяет отправить документы по уже существующим сценарию импорта и сценарию обработки пакетов непосредственно из модуля Валидации.



Необходимо выбрать Класс пакета из выпадающего списка и загрузить документы для обработки. Имя пакета формируется автоматически, согласно настройкам выбранного проекта.

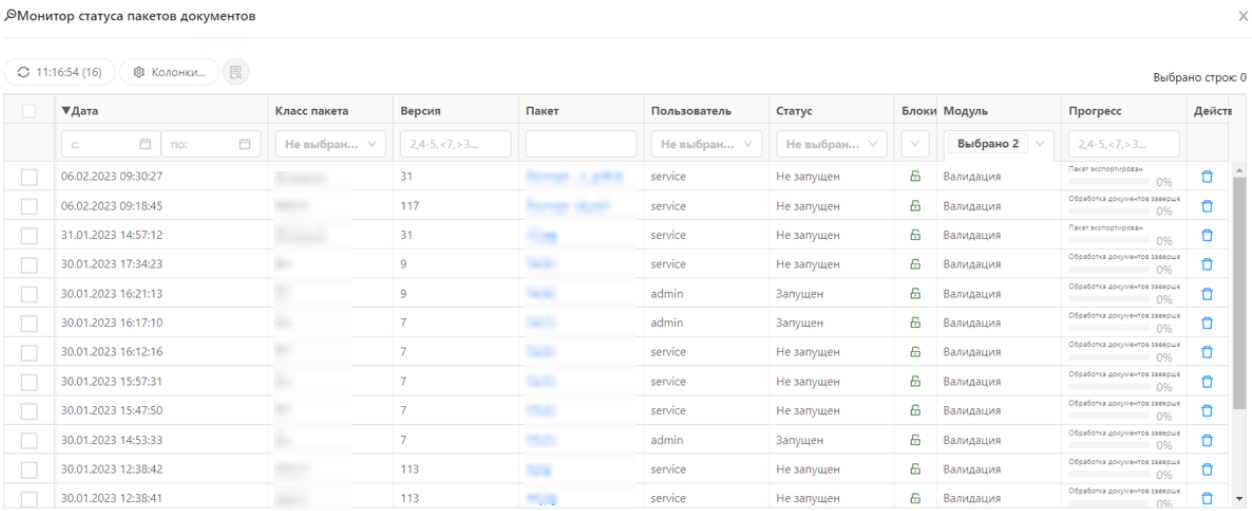


3.4. Функционал кнопки «Монитор статуса пакетов»

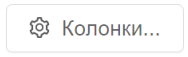
«Монитор статуса пакетов»→ 

«Монитор статуса пакетов» предназначен для отслеживания пользователем статуса пакета и выбора пакета на проверку.

Данная кнопка открывает список пакетов документов, для отслеживания их статуса в процентах.



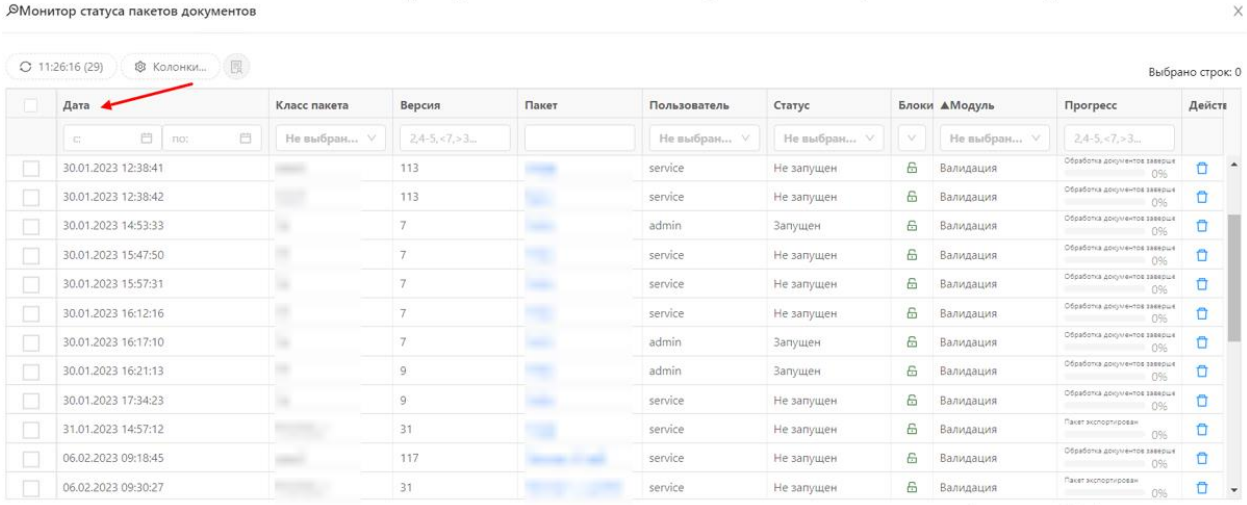
В левом верхнем углу расположена кнопка для обновления данных в списке→ 
(данная кнопка необходима для обновления строки прогресса данных)



- Выбор отображаемых столбцов, по которым можно производить фильтрацию списка пакетов.

К данному списку можно применять различные фильтры при помощи оглавления столбцов.

В столбце «Дата» можно отсортировать пакеты документов за указанный период.



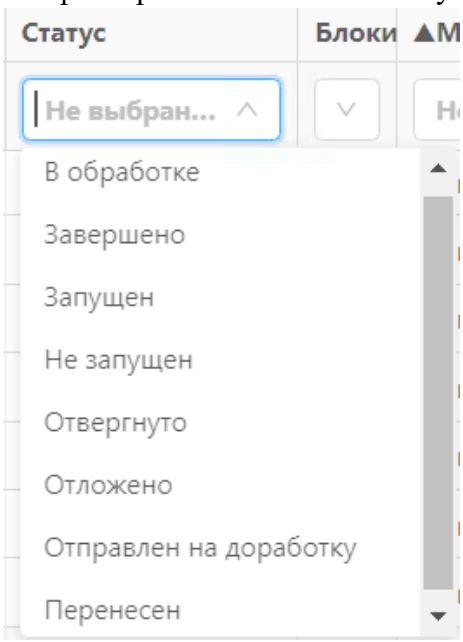
В столбце «Класс пакета» можно отфильтровать пакеты по выбранному классу пакетов.

В столбце «Версия» производится фильтрация по номеру экземпляра Класса пакета (проекта). Можно задать номер или диапазон (больше, меньше, фиксированный диапазон).

В столбце «Пакет» осуществляется поиск по имени пакета.

Столбец «Пользователь» позволяет сортировать пакеты по имени пользователя, который запустил пакеты на обработку.

В столбце «Статус» можно отфильтровать пакеты по статусу.



Напротив каждого пакета документов стоят изображения  и  .

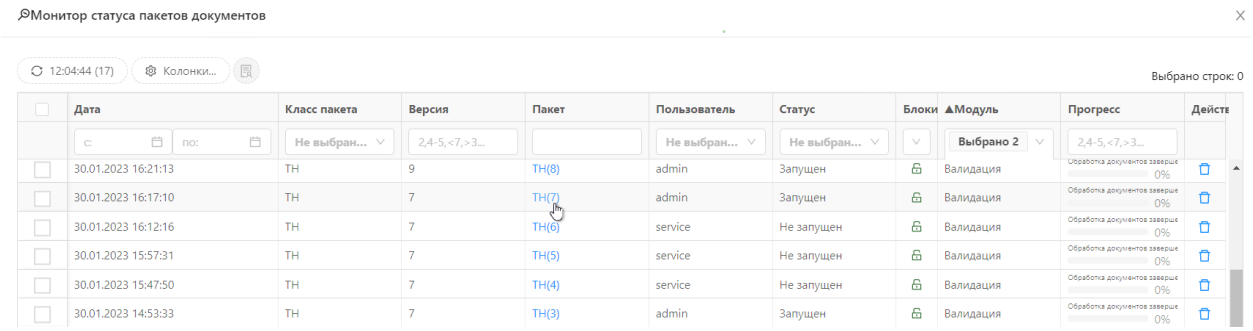
обозначает, что пакет документов открыт на Валидации кем-либо и является заблокированным для других пользователей (они не смогут его открыть).

обозначает, что с данным пакетом документов никто не работает.

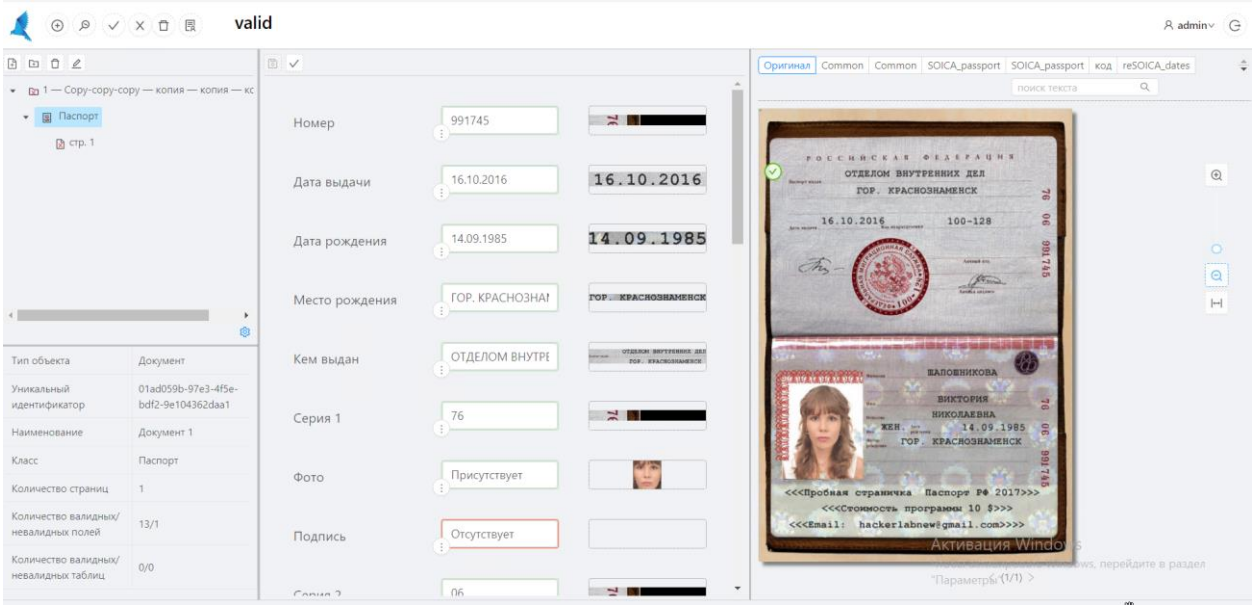
Можно удалить один или несколько пакетов из списка.

3.5. Работа с пакетом документов.

При нажатии на название любого пакета документов из списка



Откроется выбранный пакет документов для его просмотра и редактирования



В левом верхнем углу отображаются вкладки с документами, которые находятся в данном «пакете документов». На вкладке отображается название класса, присвоенного при классификации.

Над ним расположены 4 кнопки с различным функционалом.

 - Добавить документ.

 - Удалить документ.

 - Добавить страницу в документ.

Рядом с функциональными кнопками добавляются кнопки:

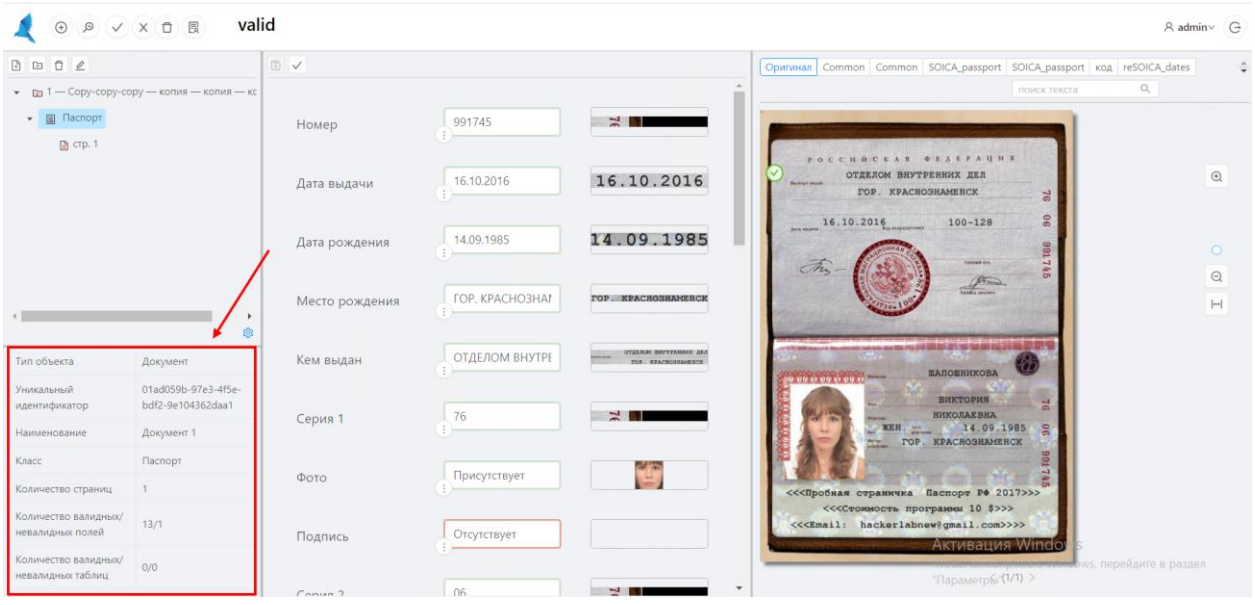
 - Принять пакет документов.

 - Закрыть пакет.

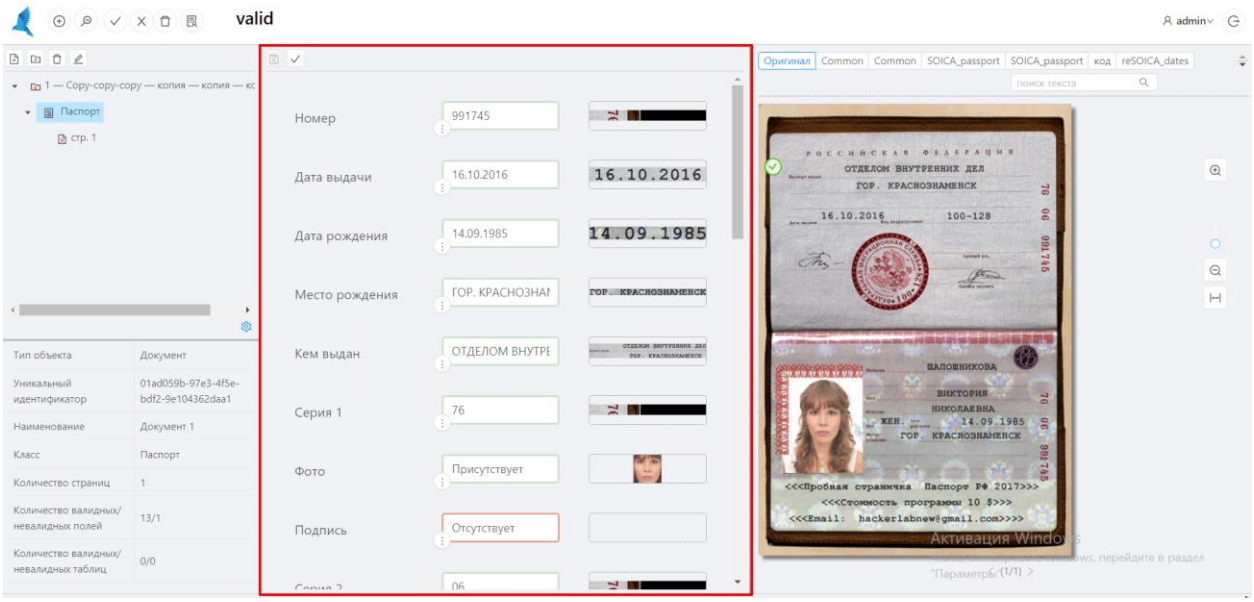
 - Удалить пакет.

 - Передать пакет пользователю или группе. Если данный оператор не имеет право принимать решение вносить данные правки в платформе Soica реализованы ролевые права. Оператор может передать пакет пользователю или группу с помощью этой кнопки. Необходимо выбрать пользователя или группу пользователей из списка.

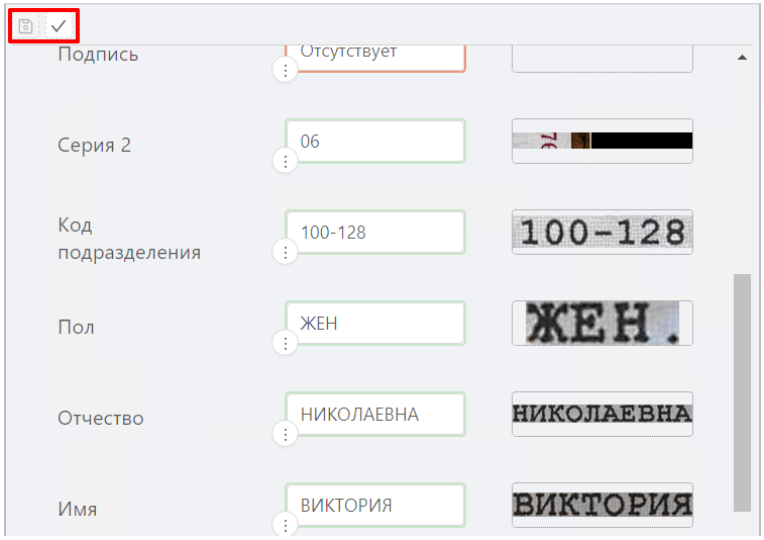
В левом нижнем углу экрана отображаются вся информация по открытому пакету.



В центральной части экрана расположены найденные данные на документе. В области представления данных выводится название поля, данные с документа и графическое отображение области поиска данных на документе (мини репрезентация), кнопки переключения между вкладками с полученными данными.



В верху центральной части расположены функциональные кнопки.

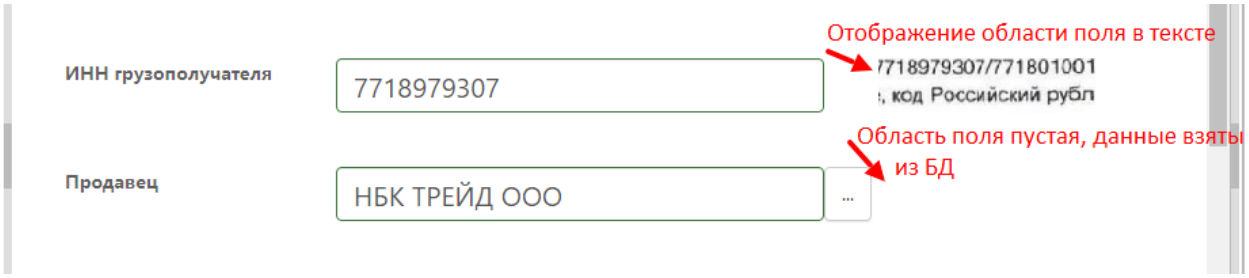


 → сохранить изменения

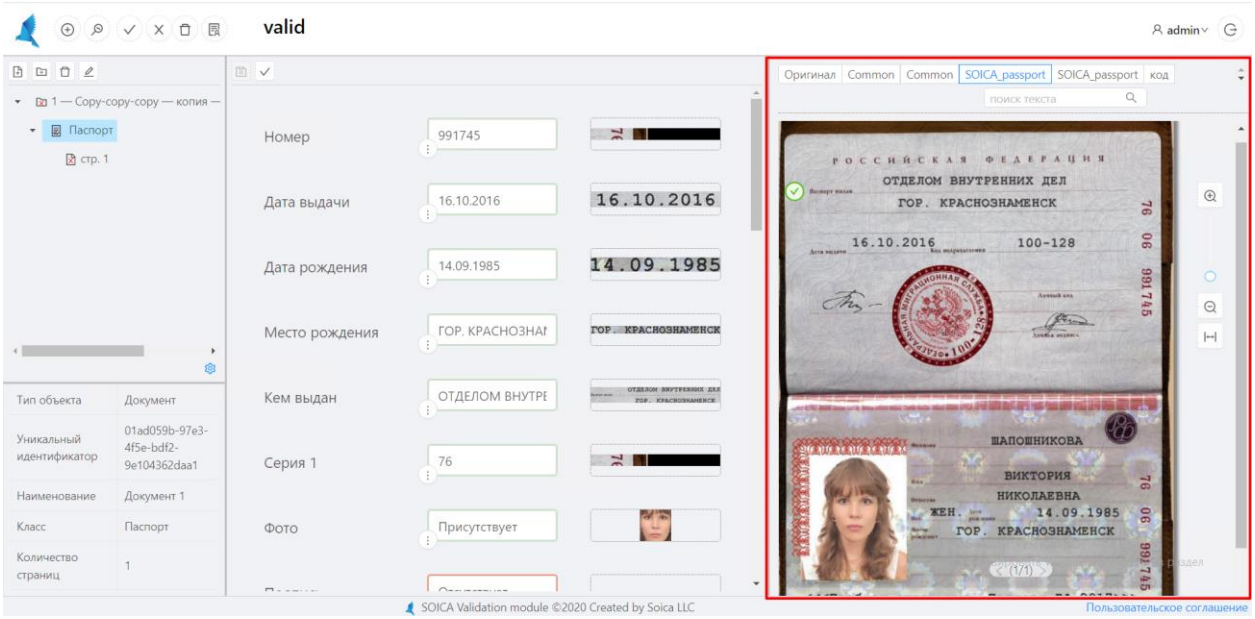
 → подтвердить всё

Наименования вкладок, их количество, распределение найденных данных между ними настраивается в модуле Администратор.

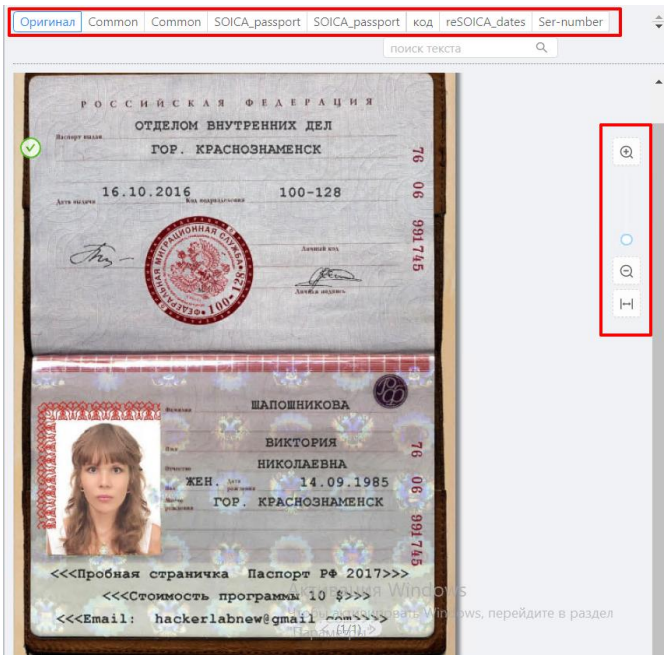
Если отображение области рядом с полем пустое, значит данное поле было заполнено из Базы данных.



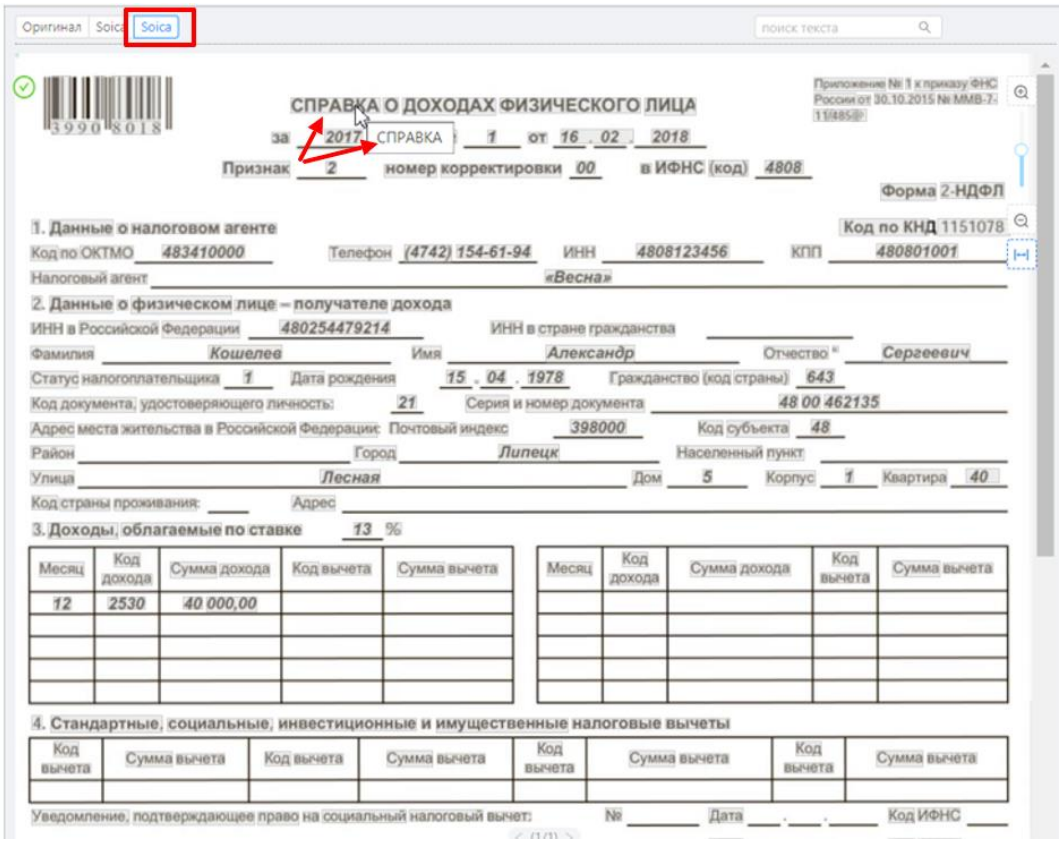
В правой части экрана расположено изображение страницы выбранного документа.

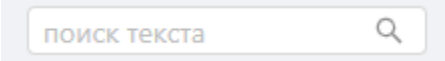


В верхней части изображения документа находятся кнопки с профилями распознавания (это те профили, которые были использованы для нахождения данных на выбранном документе). В модуле Валидации переключение между данными профилями не влияет на сам проект. Справа находятся кнопки изменения масштаба изображения документа.



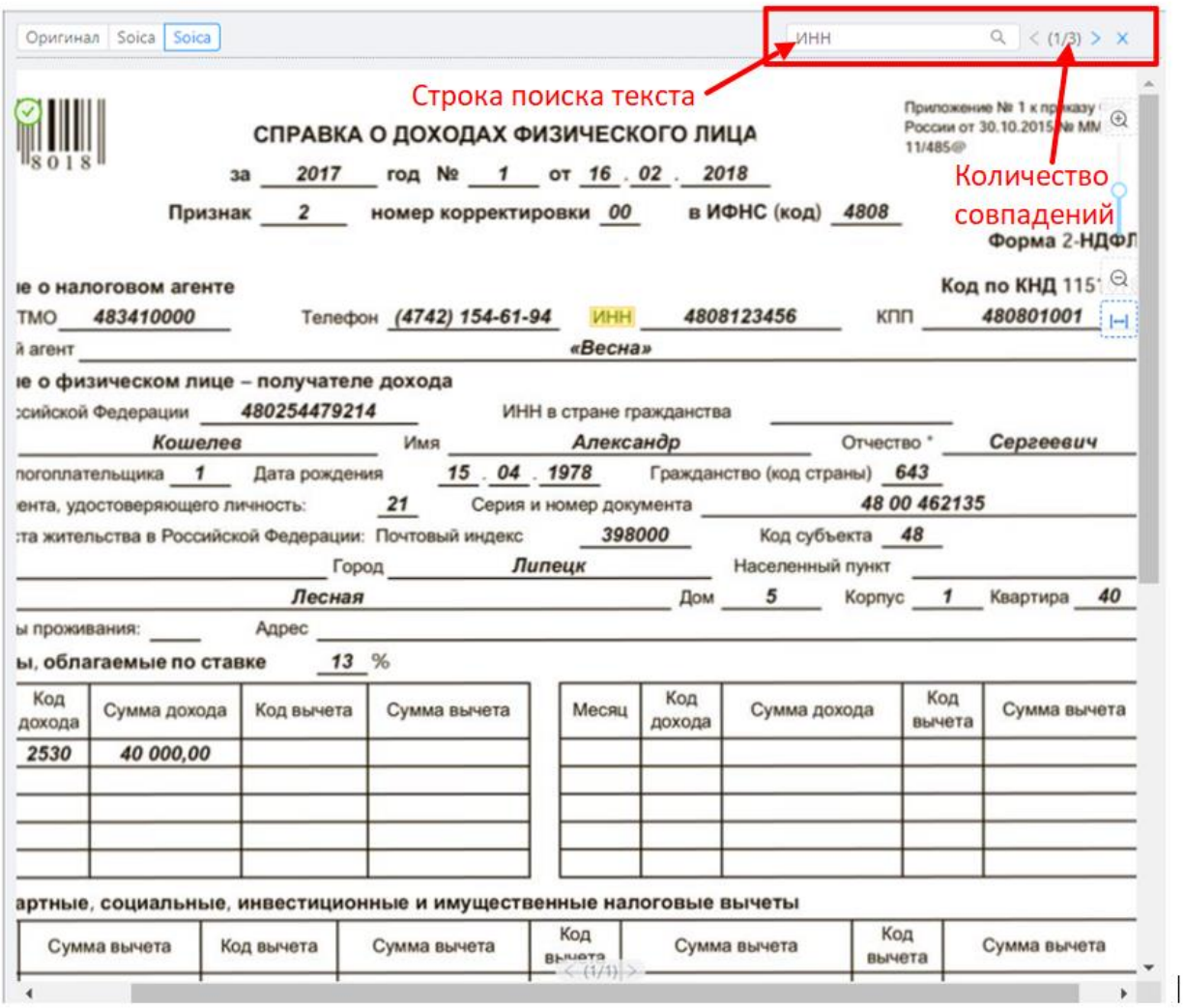
При зажатой клавише «Shift» серым подсвечиваются результаты OCR выбранного профиля распознавания. При наведении на слово отображается его OCR.



Справа от названий профилей изображения документа находится строка «Поиск текста». 

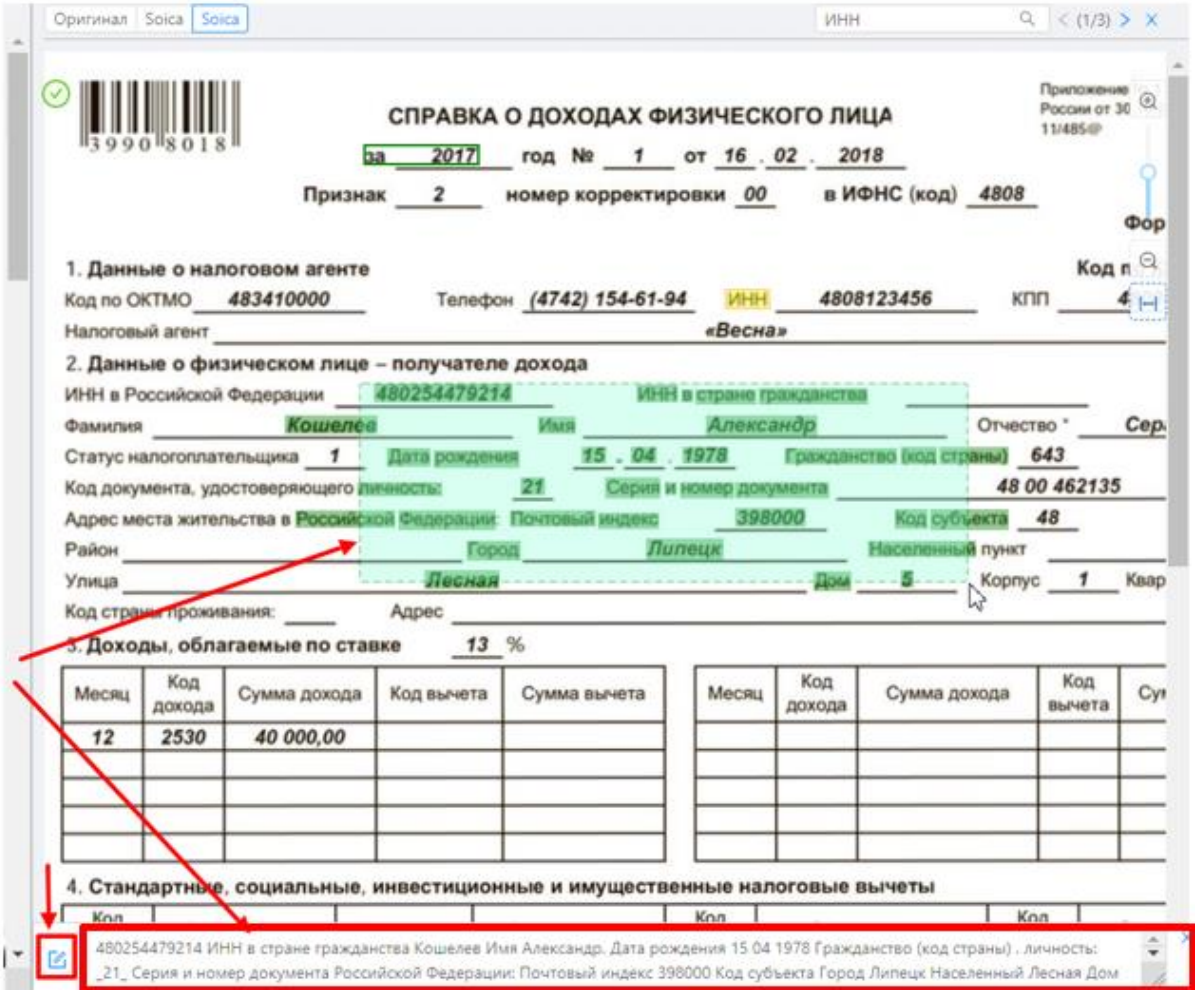
После ввода искомого текста справа от лупы отображается количество совпадений по тексту документа 

. С помощью нажатия стрелок осуществляется переход между ними. Найденный текст подсвечивается желтым цветом.



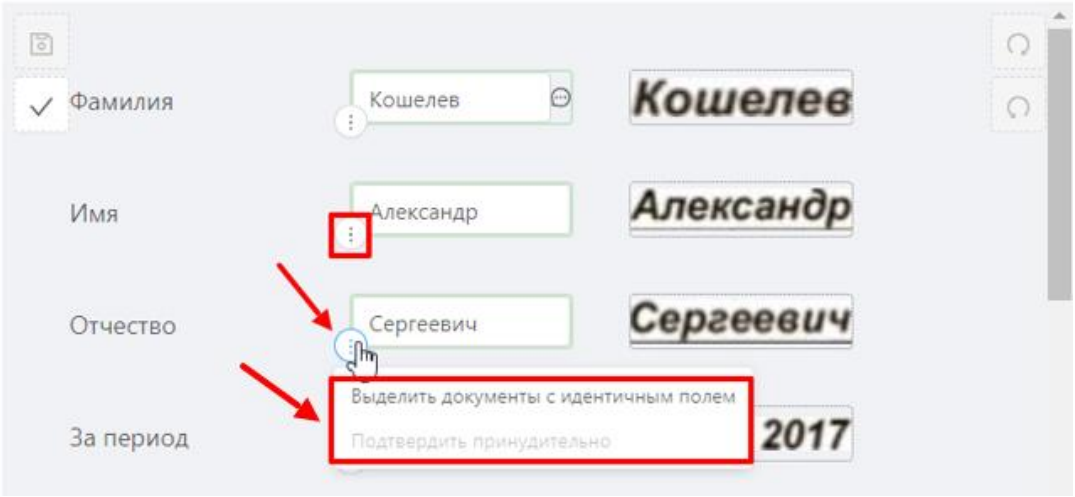
Есть возможность внести данные в нужное поле тремя способами: вручную, с помощью лассо и выбором необходимого результата OCR на самом документе.

Для ввода данных с помощью лассо необходимо установить курсор в нужном поле, далее на изображение страницы выбранного документа выбрать необходимый профилями распознавания, затем удерживая левую кнопку мыши выделить нужную зону. В окне, под изображением страницы, появится результаты OCR, попавшие в выделенную область. Их можно редактировать вручную. Чтобы добавить полученные данные в выбранное поле, необходимо нажать на кнопку  слева от окна.

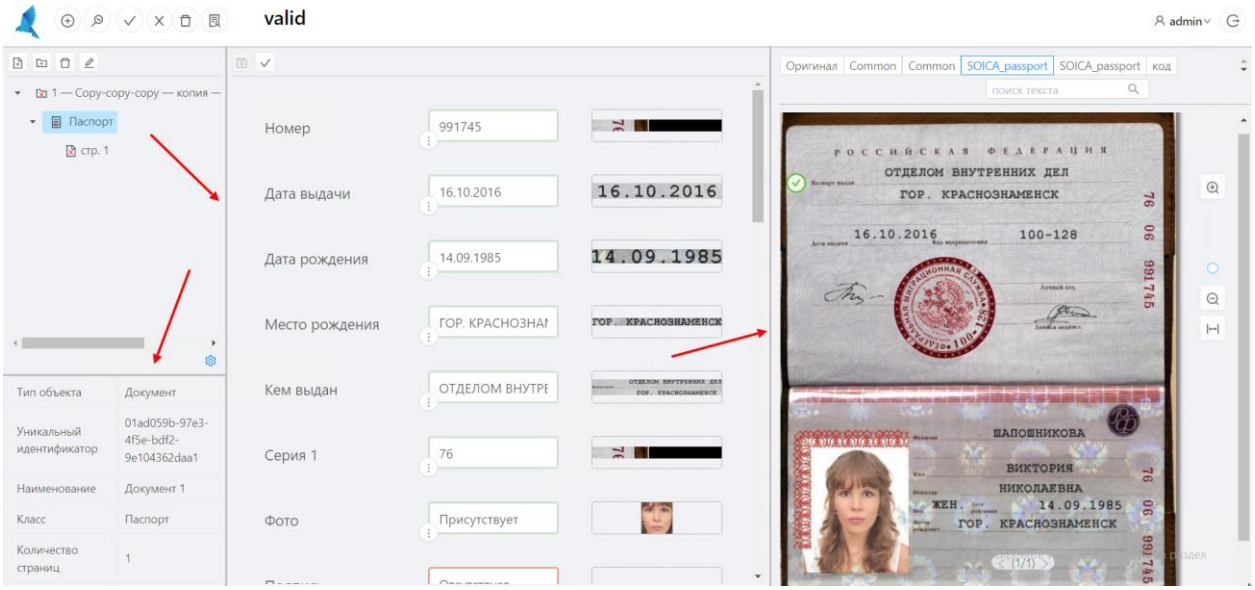


Выбором необходимого результата OCR на самом документе осуществляется аналогичным способом. На изображении необходимо выбрать необходимый профиль распознавания, далее на изображении левым щелчком мыши по нужному слову (выбранные слова подсвечиваются зеленым цветом) производится набор текста в окно под изображением. Далее нажатием клавиш «Ctrl»+«Enter» или на кнопку слева от окна происходит добавление данных в поле.

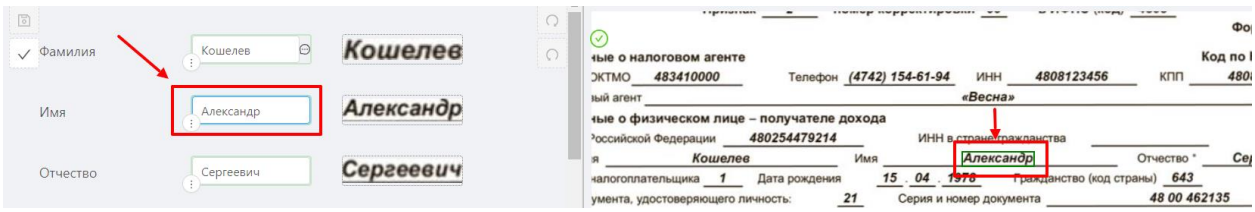
С каждым полем можно выполнить следующие действия: выделить документ с идентичным полем и подтвердить принудительно.



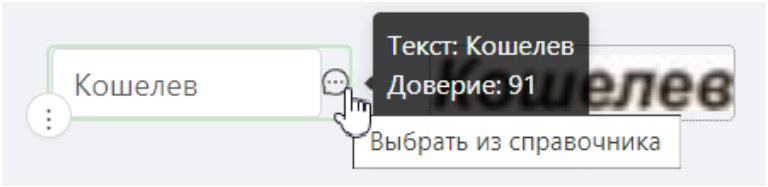
При необходимости пользователь может настроить ширину всех перечисленных выше окон. Потянув за линию в месте, указанном стрелкой.



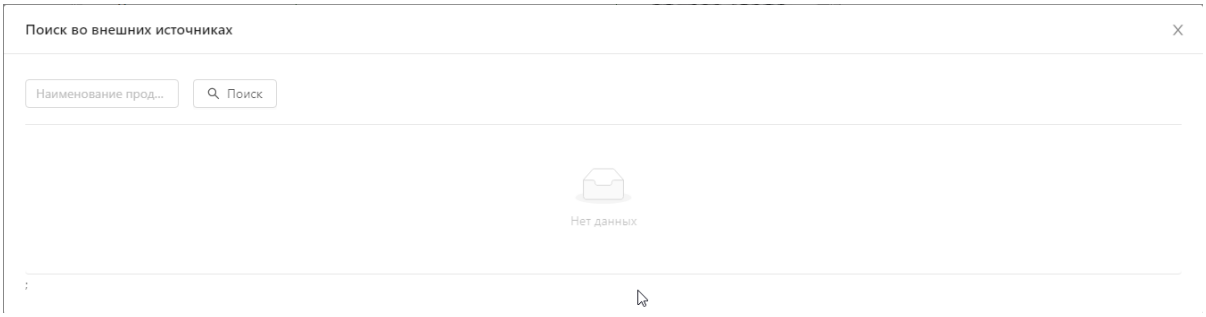
При выборе любого поля (кроме полей, которые получены из БД) на изображении документа зеленой рамкой будет подсвечена область откуда были получены данные для этого поля



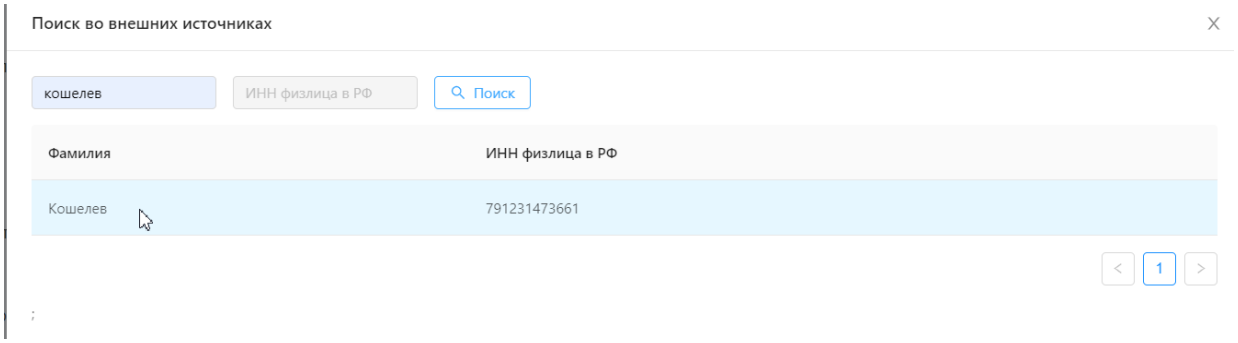
Напротив некоторых полей располагается кнопка валидации  , она предназначена для обращения и поиска в БД необходимых данных.



При нажатии на данную кнопку откроется дополнительное окно поиска в базе

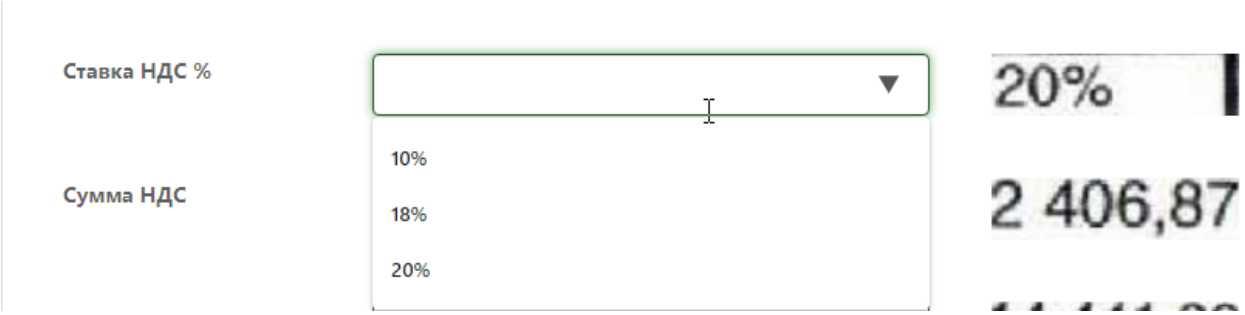


В него необходимо ввести ключевые слова (наименования/ФИО/ и т.д.) по которым будет производиться поиск в БД. После того, как данные будут внесены нажать кнопку «Поиск» и в появившемся списке выбрать необходимый результат.

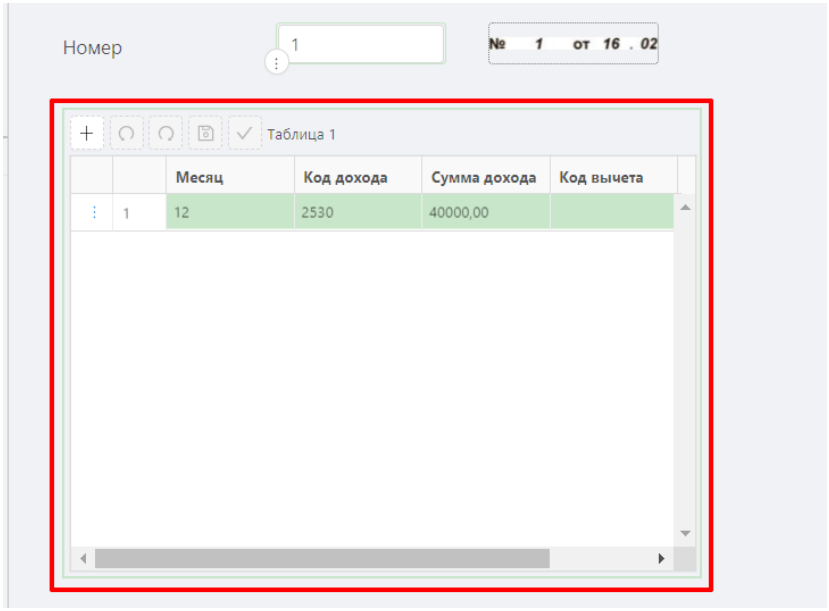


Выбранный результат отобразиться в поле. Если в базе несколько полей соответствующих полям на карточке документа, то все они будут заполнены из базы.

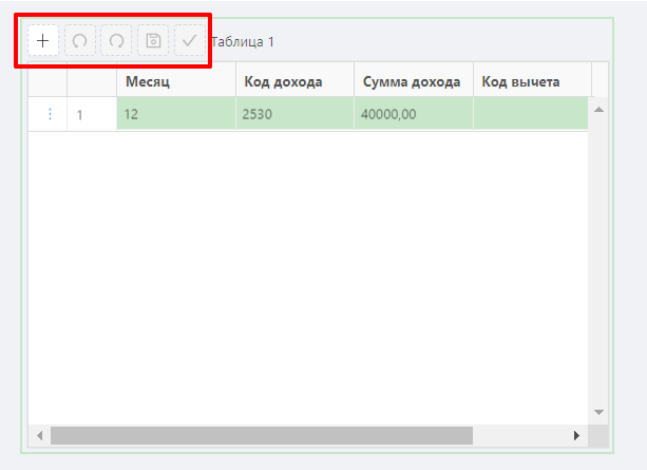
Так же в некоторых полях есть возможность выбрать необходимые данные из выпадающего списка **(для того, чтобы воспользоваться списком в данных полях необходимо удалить из данного поля все данные и когда поле будет пустым нажать на него левой кнопкой мыши)**



В модуле Валидации на карточке документа по мимо полей, отображаются найденные таблицы.

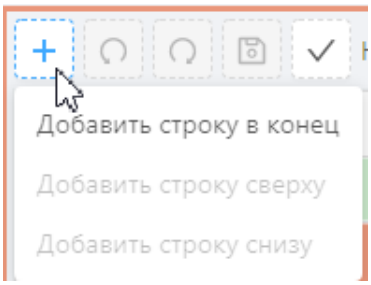


В верхнем левом углу таблицы расположены кнопки с различным функционалом



Рассмотрим каждую из них по отдельности:

 → отвечает за возможность добавления дополнительной строки в таблицу. Можно добавить новую строку в конец таблицы, сверху или снизу от выбранной строки.



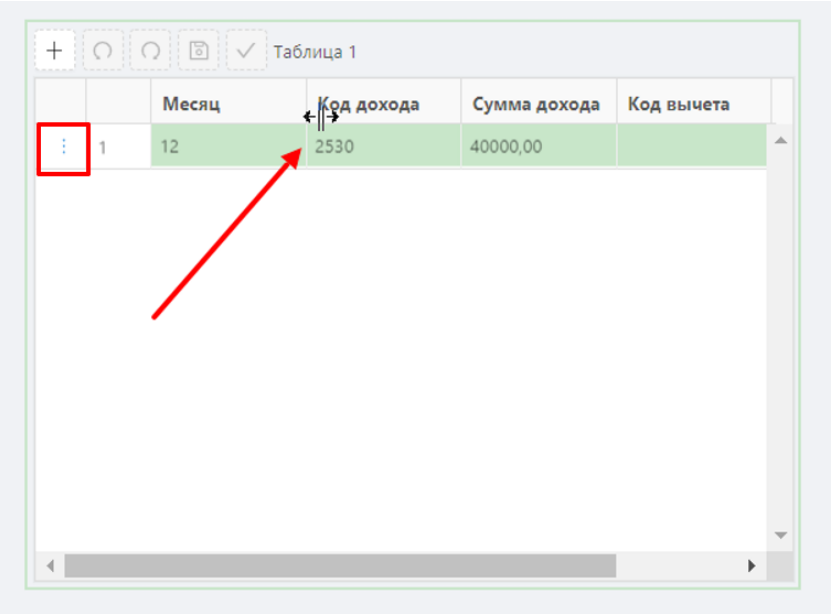
 → дает возможность отменить последнее действие (возможно использовать несколько раз подряд)

 → дает возможность вернуть последнее действие, которое было отменено (возможно использовать несколько раз подряд)

 → сохранение всех внесённых исправлений в таблицу

 → подтвердить все

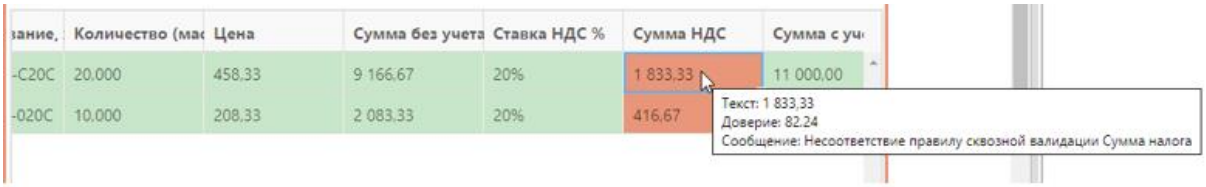
Ширину столбцов в таблицах можно редактировать.



Пользователь так же имеет возможность выполнять манипуляции со строками, нажав кнопку  около нужной строки.

Зеленым цветом в таблицах будут выделены заполненные и имеющие высокий процент доверия при распознании данные.

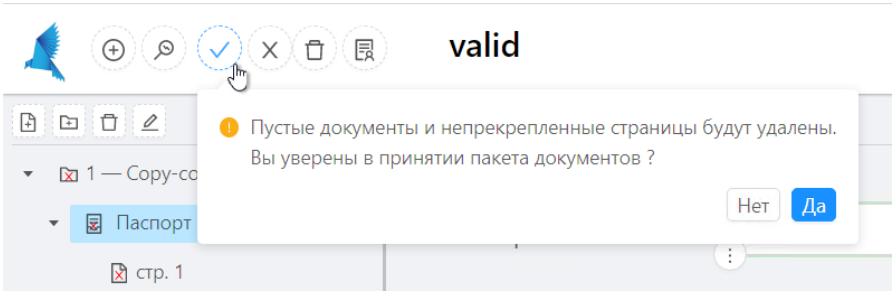
Красным цветом будут выделены не валидные ячейки. Правила валидации настраиваются в модуле Администратора. Можно настроить ячейку на определенный формат данных, на проверку наличия данных в ячейке, а также у каждой ячейки проверяется степень доверия.



При двойном нажатии левой кнопкой мыши на любую ячейку в таблице появляется возможность ее редактирования (внесения собственных данных).

Когда пользователь проверил весь документ готов передать его на Экспортирование ему необходимо нажать кнопку «Принять пакет документов»  в верхнем левом углу.

После нажатия появится сообщение



Пользователь подтверждает свое решение о принятии пакета нажатием кнопки «Да», документ отправляется на Экспорт и пропадает из модуля Валидации.

4. Экспорт.

Модуль экспорта предназначен для экспорта полученных данных.

Экспортирование – процесс вывода найденных данных в заданном виде. Количество сценариев экспорта не ограничено. Каждый документ может быть экспортирован по разным маршрутам и в разных форматах параллельно. Настройка сценариев ведется в модуле администратора.

В системе предусмотрена возможность сохранения в сетевом каталоге, возможность сохранения данных в локальном каталоге.

При отправке результатов по электронной почте адресат может быть задан заранее или сформирован на основе данных находящихся на распознанных документах.

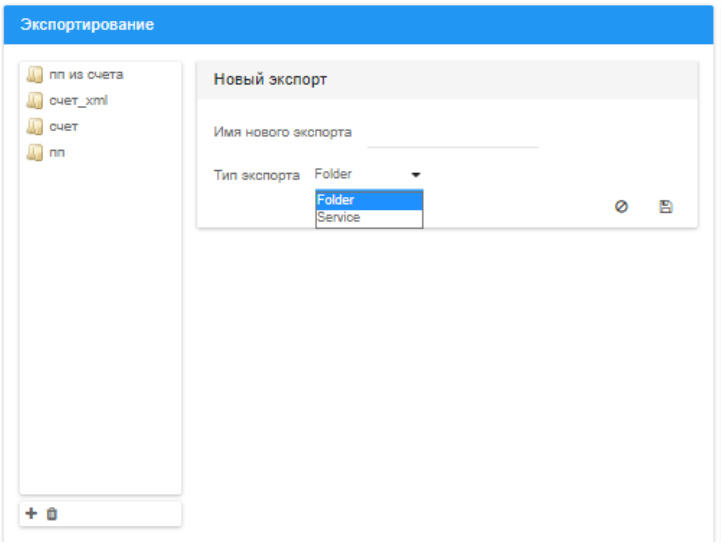
В качестве адресата может выступать и отправитель если документы пришли в SOICA по электронной почте.

Результатом экспорта будет 3 файла на каждый обработанный пакет. 1 файл выбранного формата с данными (xml, txt ,docx, xlsx), 1 файл xml и один файл с исходным изображением.

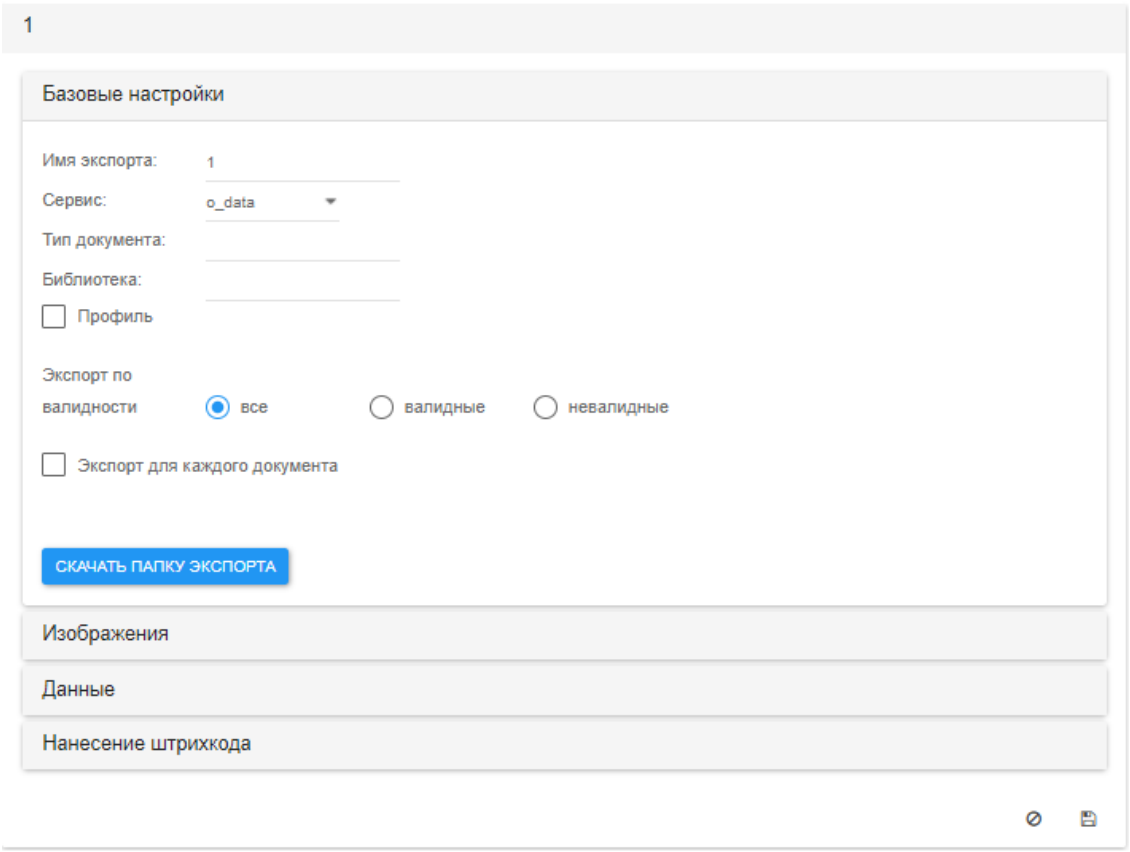
Область экспортирования делится на два раздела: Список сценариев экспорта и Настройка выбранного сценария.

Для создания нового задания экспорта необходимо нажать + и в открывшемся окне ввести имя задания экспорта. Так же необходимо выбрать тип экспорта в папку (Folder) или в сервис (Service).

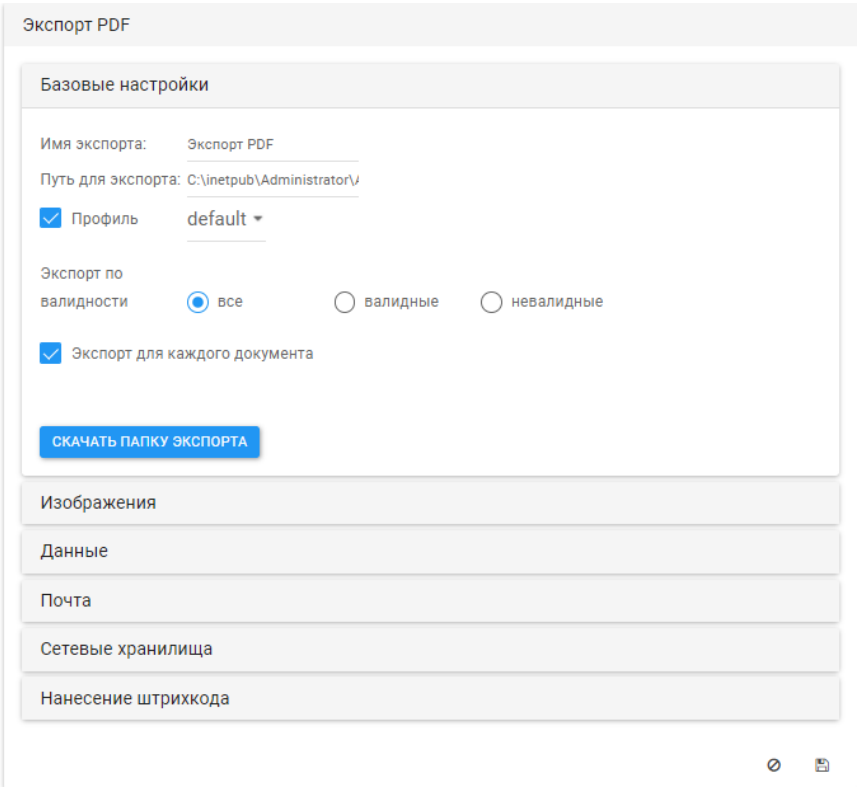
При выборе варианта в сервис, в базовых настройках из выпадающего списка необходимо выбрать сервис, заранее добавленный в веб сервисах.



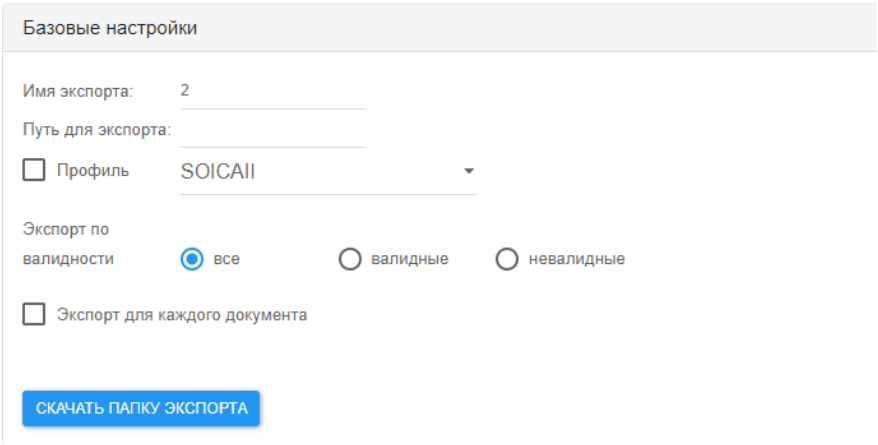
(Рис. 114 Выбор типа экспорта при создании нового сценария)



(Рис. 114.1 Интерфейс настройки экспорта. Тип экспорта Service)



(Рис. 114.2 Интерфейс настройки экспорта. Тип экспорта Folder)



(Рис. 115 Экспорт в папку – Базовые настройки)

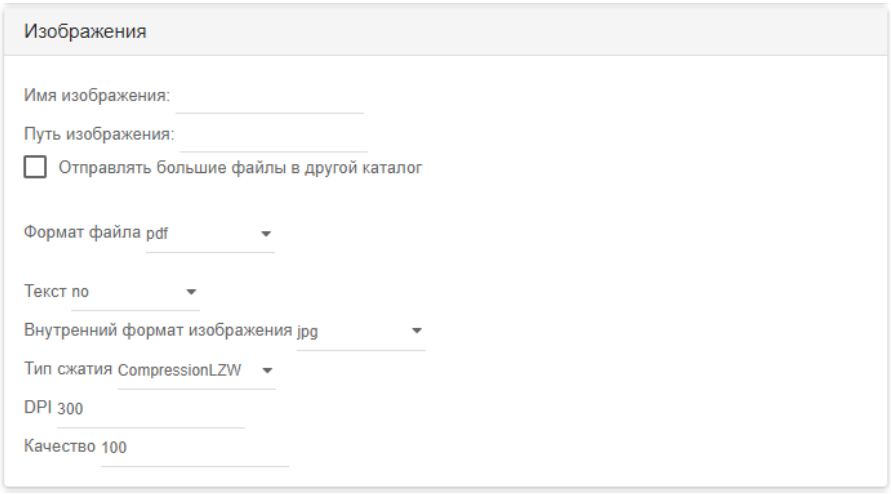
Путь для экспорта. Указывает директорию, в которую будут помещаться экспортируемые файлы. Выбрать из имеющихся, либо создать новую (...)

Профиль. Указывает имя профиля распознавания, репрезентации страниц полученные которым будут экспортироваться.

Экспорт по валидности. Необходимо выбрать нужное условие для экспортирования: только валидные документы, все или любые.

Экспорт для каждого документа. Эта опция влияет на экспортирование всех документов. Будут ли все документы экспортироваться в один файл для изображения и данных для пакета или в разные.

Скачать папку с данными. Кнопка позволяет скачать с сервера папку экспорта в виде архива.



(Рис. 116 Экспорт – Изображение)

Имя изображения. Указывается имя, которое автоматически сформируется при экспорте.

Путь изображения, Путь данных. Путь указывается относительно корневого каталога экспорта. Поле может содержать переменные типа содержимого системных значений пакета и документа, значений полей пакета и документа. Например: путь для экспорта: «C:\export», путь данных: «df:Номер_накладной». Итоговый путь может выглядеть так: «C:\export\TH003454654\».Указывается если нужно положить изображение в отдельную папку.

Имя данных. Указывается имя, которое автоматически сформируется при экспорте.

Конструктор имен и путей к файлам. Указывает строку, состоящую из констант, имен полей, или системных значений по которой формируется конечная строка, определяющая (разделенных символом «|»): имя и путь к файлам изображения и данных, получателя, тему и тело письма (для отправки файлов почтой). В качестве деталей для конструктора могут быть: Константа, Имя поля документа, Имя поля пакета, Системное значение документа, Системное значение пакета.

Константа. Указывает строку, составляющую часть итоговой строки конструктора.

Имя поля документа. Указывается в формате: df:Поле. При формировании строки, этот текст заменяется на текст указанного поля.

Имя поля пакета. Указывается в формате: bf:Поле. При формировании строки, этот текст заменяется на текст указанного поля.

Системное значение документа. Указывается в формате: xdc:Значение. При формировании строки, этот текст заменяется на указанное значение. Варианты значений: dc_name, d_name, page_count, d_uniq_id:

dc_name. Имя класса документа.

d_name. Имя документа.

page_count. Количество страниц.

d_uniq_id. Уникальный идентификатор

Системное значение пакета. Указывается в формате: xbc:Значение. При формировании строки, этот текст заменяется на указанное значение. Варианты значений: bc_name, date, time, user_name, pc_name, b_guid, b_name, page_count, doc_count:

bc_name. Имя класса пакета.

date. Дата. Указывает дату создания пакета.

time. Время. Указывает время создания пакета.

user_name. Имя пользователя. Указывает пользователя, создавшего пакет.

pc_name. Имя компьютера. Указывает имя компьютера, на котором был создан пакет.

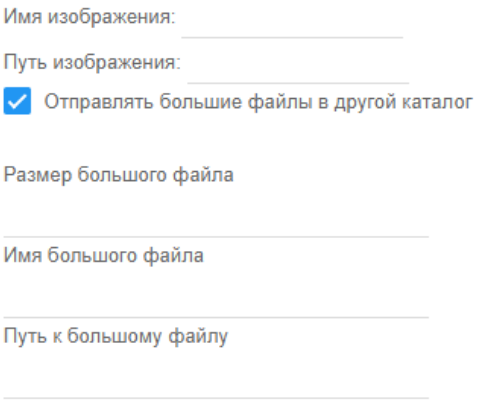
b_guid. GUID пакета.

b_name. Имя пакета.

page_count. Количество страниц.

doc_count. Количество документов.

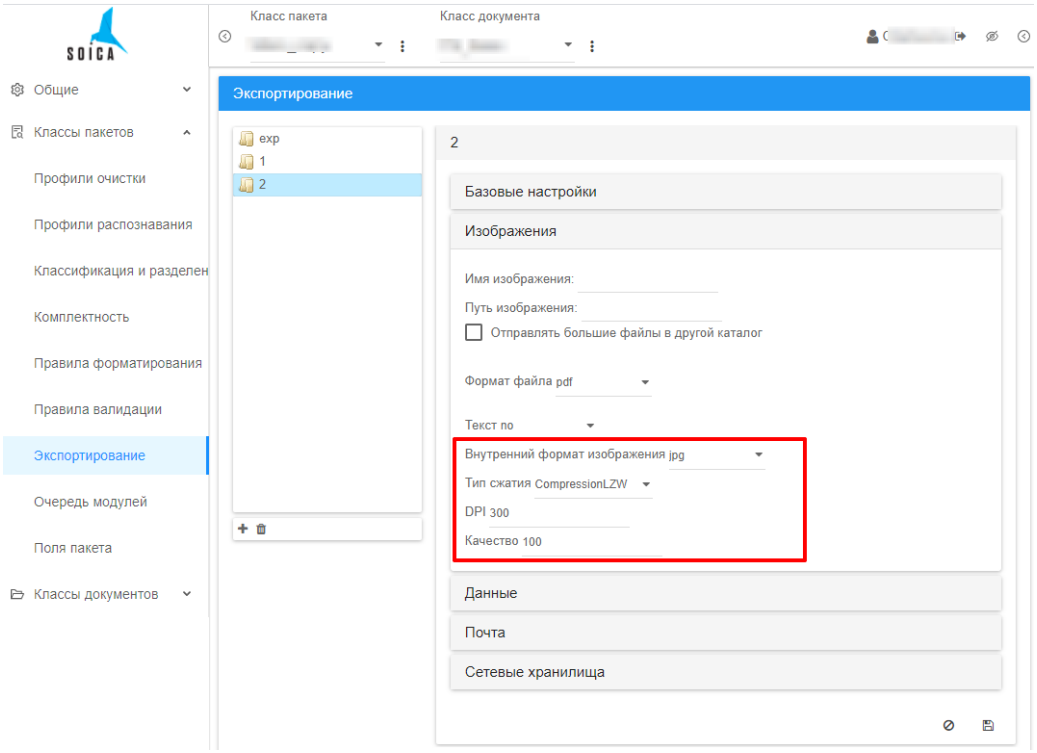
Отправлять большие файлы в другой каталог. При выборе данной опции необходимо указать размер большого файла, Имя большого файла и Путь к большому файлу.



Формат вывода изображения. Указывает формат для вывода изображения документов. Варианты: pdf, tiff, none (не выгружать изображение), jpg.

Также необходимо выбрать параметры создания выгружаемого файла:

- Внутренний формат изображения: jpg, tiff, png
- Тип сжатия DPI
- Качество изображения



При помощи вышеуказанных параметров можно изменить формат и размеры исходных файлов для формирования выгружаемого файла и уменьшить его размер.

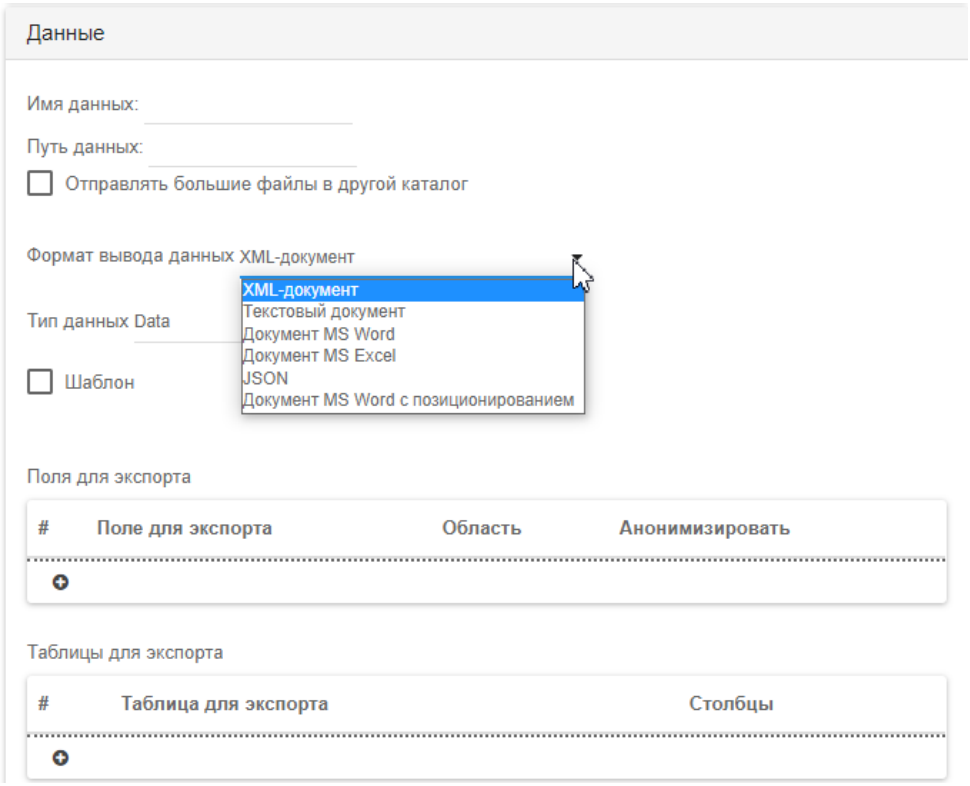
Вывод изображения в формате PDF. В данном случае из страниц документа будет сформирован pdf файл.

Добавление текста в pdf файл. В экспортируемый pdf документ можно добавить текст. Варианты: не добавлять, добавить OCR из репрезентации, добавить поля и таблицы.

Добавить результаты OCR. В документ будет добавлен текст из результатов OCR указанной репрезентации.

Вывод изображения в формате tiff. В данном случае из страниц документа будет сформирован многостраничный tiff файл.

Сжатие tif. Указывает тип сжатия для tif файла. Варианты: "Сжатие LZW", "Сжатие CCITT3", "Сжатие CCITT4", "Сжатие Rle", "Без сжатия".



(Рис. 117 Экспорт – Данные)

Формат вывода данных. Указывает тип файла для записи в него информации о извлеченных полях и таблицах документа. Варианты: "XML-документ", "Текстовый документ", "Документ MS Word", "Документ MS Excel".

- **XML-документ**. Вывод данных в формате XML, файл *.xml.
- **Текстовый документ**. Вывод данных в текстовом формате, файл *.txt.
- **Документ MS Word**. Вывод данных в формате документа MS Word, файл *.docx.
- **Документ MS Excel**. Вывод данных в формате документа MS Excel, файл *.xlsx.
- **JSON.** Вывод данных в формате документа JSON, файл *.json.
- **Документ MS Word с позиционированием.** Вывод данных в формате документа MS Word, файл *.docx. Каждый обзац помещается в таблицу без границ, с позиционированием, как на исходном документе.

Красить ячейки. Опция доступна при выборе формата вывода данных Документ MS Excel. Позволяет настроить цветовой вид создаваемого файла в соответствии с ячейками данных из валидных и не валидных полей документа. Ячейки с данными из валидных полей документа будут иметь заливку зеленого цвета, а из не валидных – красного.

Тип данных. Указывает что будет записываться в качестве данных. Варианты: поля и таблицы, результаты OCR.

Вывод полей и таблиц. В данном случае будет произведена запись текста полей и таблиц в выходной файл.

Поля для экспорта. Содержит список имен полей, текст которых необходимо экспортировать.

Экспортировать область поля. Если выбрана данная опция, то рядом с файлом, содержащим данные, будут сохраняться изображения областей полей.

Таблицы для экспорта. Содержит список таблиц, текст ячеек которых необходимо экспортировать.

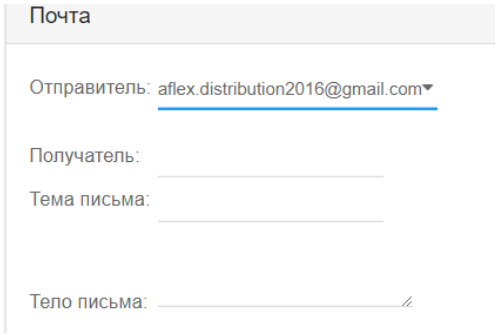
Экспортируемые столбцы таблицы. Содержит список номеров столбцов таблицы, текст ячеек которых будет экспортирован.

Шаблон. Если данная опция выбрана, то данные будут экспортироваться в соответствии с указанным шаблоном. Шаблон должен быть формата: для xml и txt – xslt схемы, для docx – docx, для xlsx – xlsx.

Путь к файлу шаблона. Указывает путь к файлу, который будет заполняться экспортируемыми данными.

Список сопоставления закладок из шаблона с полями и таблицами документа. Содержит пары: имя закладки в шаблоне, имя поля/таблицы в документе.

Если указаны типы экспорта данных, то сопоставления закладок и полей реализованы внутри файла шаблона, которым являются xslt схемы. При экспорте docx, в качестве закладок используются элементы «Закладка» (bookmark). При экспорте xlsx, указываются адреса ячеек.



(Рис. 118 Экспорт - Почта)

Отправка экспортируемых файлов по email. Если заполнены соответствующие поля, то будет произведена попытка отправки файлов с изображением и данными по электронной почте.

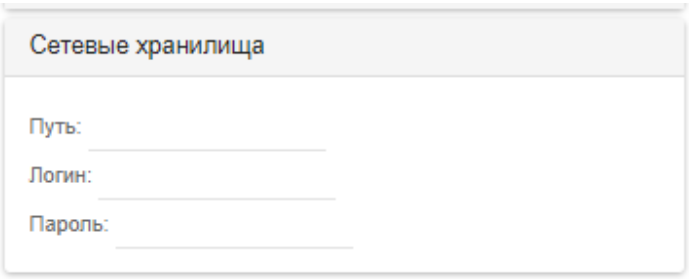
Отправитель. Указывает электронную почту отправителя письма, выбирается из заранее заданного списка.

Получатель. Указывает адрес электронной почты получателя письма.

Тема письма. Указывает тему отправляемого письма.

Тело письма. Указывает текст отправляемого письма.

Сетевые хранилища. Отправка экспортированных файлов по протоколу ftp, либо на сетевую папку, если путь не начинается с «ftp». Если заполнены соответствующие поля, то будет произведена попытка отправки файлов с изображением и данными по ftp протоколу.

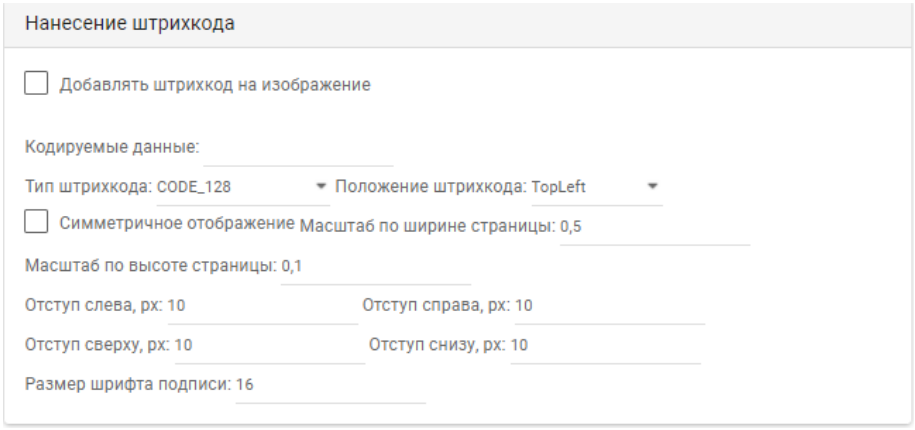


(Рис. 119 Экспорт – Сетевые хранилища)

Путь к ftp директории или на сетевую папку. Указывает полный путь к директории, на ftp сервере, либо путь к сетевой папке, куда должны поместиться экспортируемые файлы.

Логин. Логин для аутентификации на ftp сервере либо к сетевой папке.

Пароль. Пароль для аутентификации на ftp сервере либо к сетевой папке.



(Рис. 120 Экспорт – Нанесение штрих кода)

Позволяет добавлять штрих код на изображение по заданным параметрам.

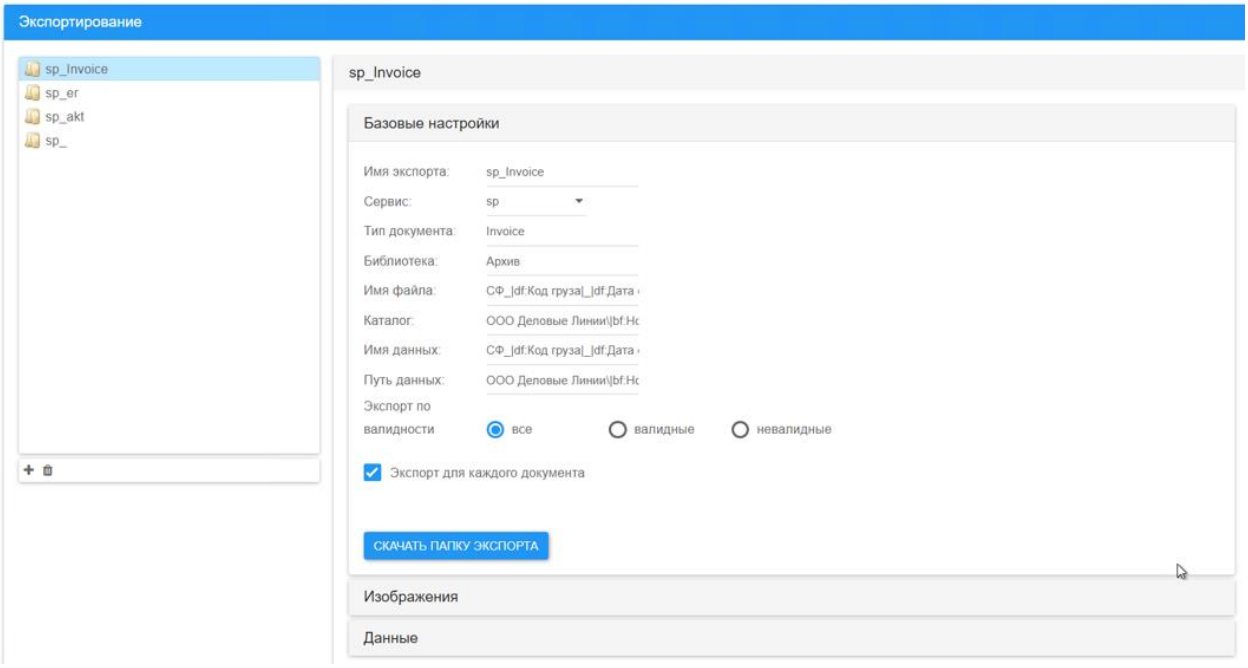
4.1. Настройки сценария экспорта в SharePoint

Сценарий экспорта SharePoint в модуле администратора

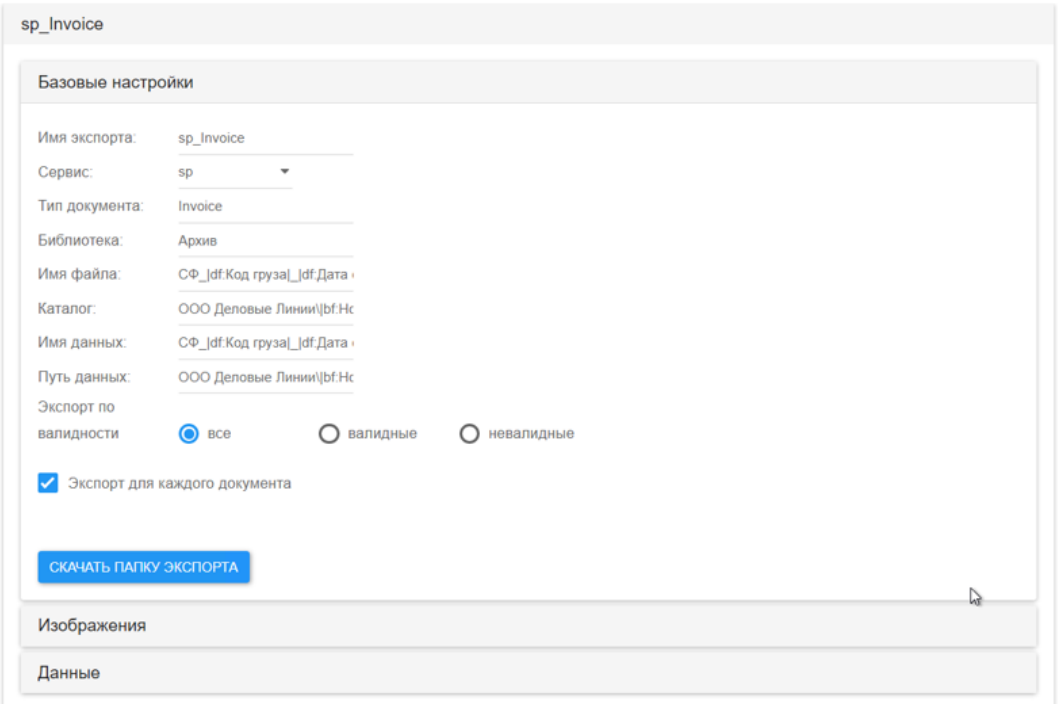
Платформа Microsoft SharePoint – это программный продукт, основной задачей которого является предоставление сотрудникам компании инструмента эффективного взаимодействия.

В програмном продукте Soica есть возможноть настроить сценарии экспорта в SharePoint.

Экспортирование – процесс вывода найденных данных в заданном виде. Количество сценариев экспорта не ограничено. Каждый документ может быть экспортирован по разным маршрутам и в разных форматах параллельно. Настройка сценариев ведется в модуле администратора.



(Рис.121 Интерфейс настройки экспорта SharePoint)



(Рис.122 Экспорт SharePoint – Базовые настройки)

Сервис. Из выпадающего списка выбирается сервис SharePoint, заранее добавленный на вкладке Веб-сервисы. К этому сервису будет обращаться сценарий экспорта для выгрузки данных в SharePoint.

Тип документа. Наименование типа контента в SharePoint

Библиотека. Наименование библиотеки в SharePoint.

Имя файла. Указывается имя файла изображения, которое автоматически сформируется при экспорте

Каталог. Путь к каталогу в библиотеке SharePoint, в который будет помещен файл с изображением. Поле может содержать переменные типа содержимого системных значений пакета и документа, значений полей пакета и документа.

Путь данных. Путь к каталогу в библиотеке SharePoint, в который будет помещен файл с данными. Поле может содержать переменные типа содержимого системных значений пакета и документа, значений полей пакета и документа. Если оставить это поле пустым, выгрузка файла с данными не будет производится.

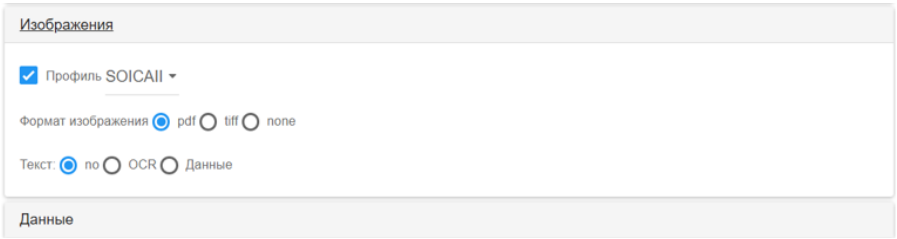
Имя данных. Указывается имя, которое автоматически сформируется при экспорте.

Экспорт по валидности. Необходимо выбрать нужное условие для экспортирования: только валидные документы, все или любые.

Экспорт для каждого документа. Эта опция влияет на экспортирование всех документов. Будут ли все документы экспортироваться в один файл для изображения и данных для пакета или в разные.

Конструктор имен и путей к файлам. Указывает строку, состоящую из констант, имен полей, или системных значений по которой формируется конечная строка, определяющая (разделенных символом «|»): имя и путь к файлам изображения и данных, получателя, тему и тело письма (для отправки файлов почтой). В качестве деталей для конструктора могут быть: Константа, Имя поля документа, Имя поля пакета, Системное значение документа, Системное значение пакета.

Константа. Указывает строку, составляющую часть итоговой строки конструктора.



(Рис. 123 Экспорт – Изображение)

Профиль. Указывает имя профиля распознавания, репрезентации страниц полученные которым будут экспортироваться.

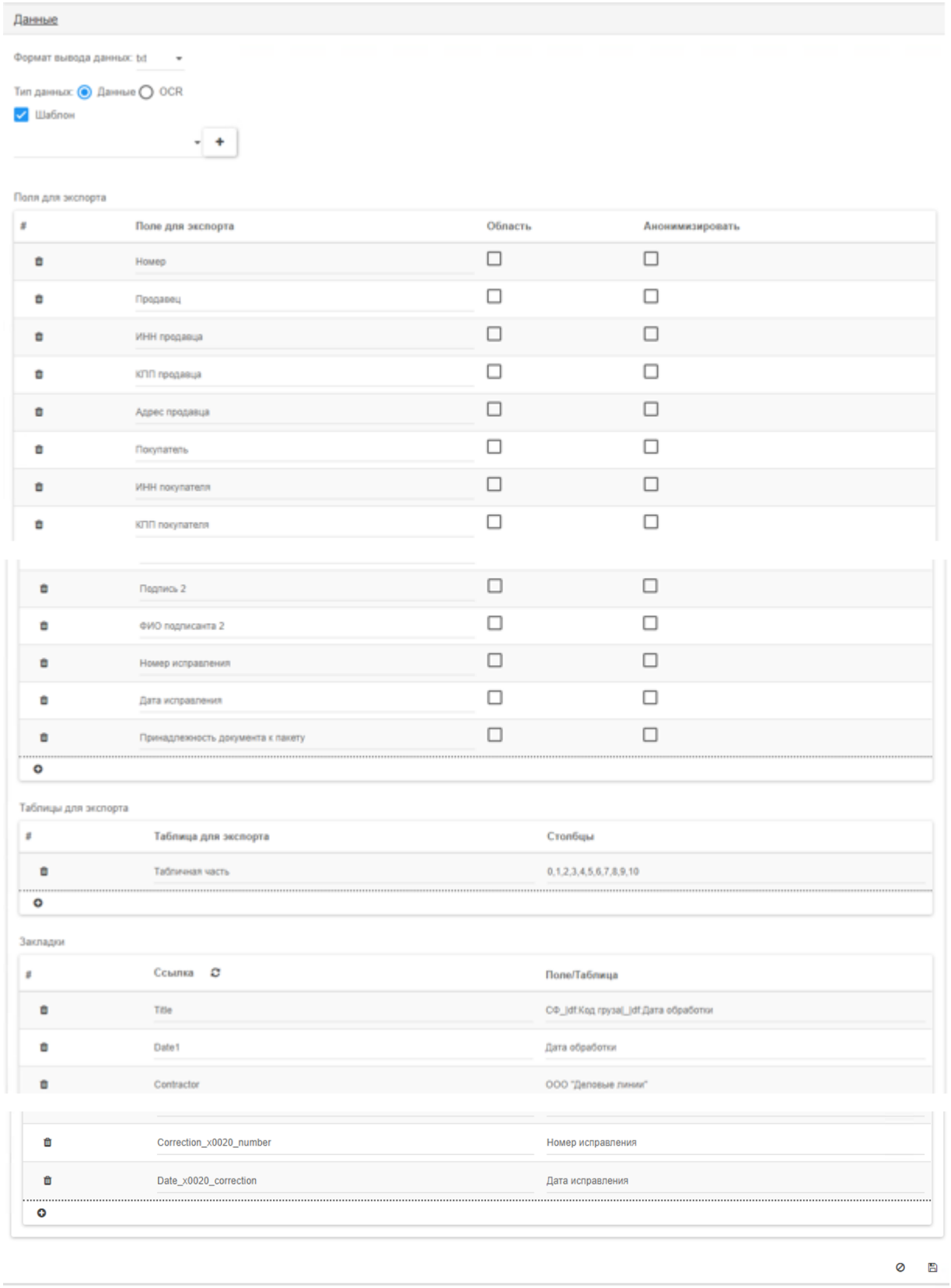
Формат вывода изображения. Указывает формат для вывода изображения документов. Варианты: pdf, tiff, none (не выгружать изображение).

Вывод изображения в формате PDF. В данном случае из страниц документа будет сформирован pdf файл.

Добавление текста в pdf файл. В экспортируемый pdf документ можно добавить текст. Варианты: не добавлять, добавить OCR из репрезентации, добавить поля и таблицы.

Добавить результаты OCR. В документ будет добавлен текст из результатов OCR указанной репрезентации.

Вывод изображения в формате tiff. В данном случае из страниц документа будет сформирован многостраничный tiff файл.



(Рис.124 Экспорт – Данные)

Формат вывода данных. Указывает тип файла для записи в него информации о извлеченных полях и таблицах документа. Варианты: "XML-документ", "Текстовый документ", "Документ MS Word", "Документ MS Excel".

- **XML-документ**. Вывод данных в формате XML, файл *.xml.
- **Текстовый документ**. Вывод данных в текстовом формате, файл *.txt.
- **Документ MS Word**. Вывод данных в формате документа MS Word, файл *.docx.
- **Документ MS Excel**. Вывод данных в формате документа MS Excel, файл *.xlsx.

Тип данных. Указывает что будет записываться в качестве данных. Варианты: поля и таблицы, результаты OCR.

Вывод полей и таблиц. В данном случае будет произведена запись текста полей и таблиц в выходной файл.

Поля для экспорта. Содержит список имен полей, текст которых необходимо экспортировать.

Экспортировать область поля. Если выбрана данная опция, то рядом с файлом, содержащим данные, будут сохраняться изображения областей полей.

Таблицы для экспорта. Содержит список таблиц, текст ячеек которых необходимо экспортировать.

Экспортируемые столбцы таблицы. Содержит список номеров столбцов таблицы, текст ячеек которых будет экспортирован.

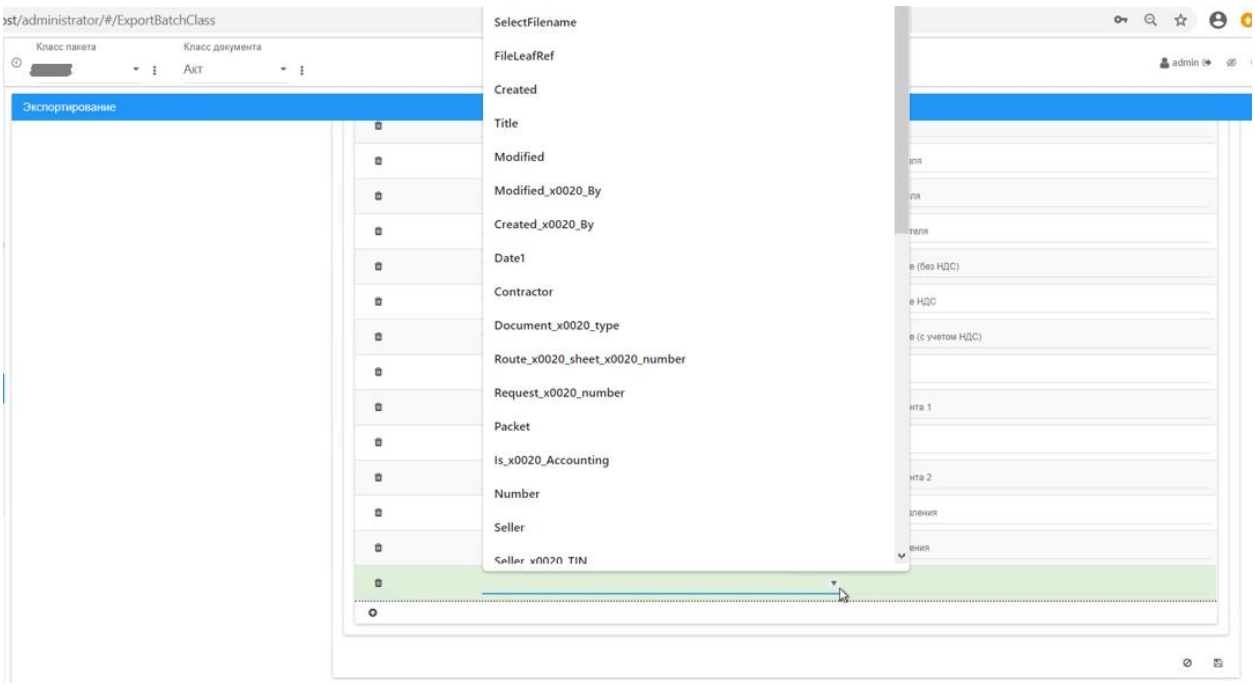
Путь к файлу шаблона. Указывает путь к файлу шаблона, с помощью которого будет формироваться файл с данными.

Список сопоставления полей из SharePoint с полями и таблицами документа. Список полей метаданных для определенного типа контента из SharePoint. Содержит пары: имя поля в SharePoint, имя поля/таблицы в документе.

Для синхронизации полей необходимо нажать кнопку  .

При смене библиотеке или типа документа необходимо заново нажать на кнопку синхронизации для актуализации полей из SharePoint.

Выбор полей производится из выпадающего списка.



(Рис. Список полей метаданных для определенного типа контента из SharePoint)

SharePoint

Созданный с помощью Microsoft SharePoint внутрикорпоративный портал становится единым коммуникационным центром и электронным хранилищем информации одновременно.

Внутренний корпоративный портал на базе Microsoft SharePoint позволяет:

- Организовать единую среду для совместной работы над документами и проектами Показать\спрятать подробности.

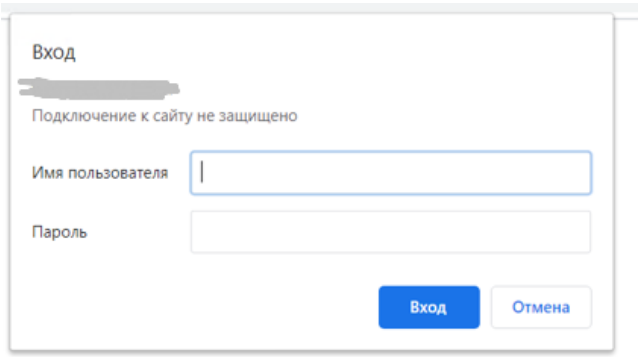
- Построить онлайн-площадку для общения по различным рабочим вопросам Показать\спрятать подробности.

- Создавать решения для автоматизации любых рабочих процессов.

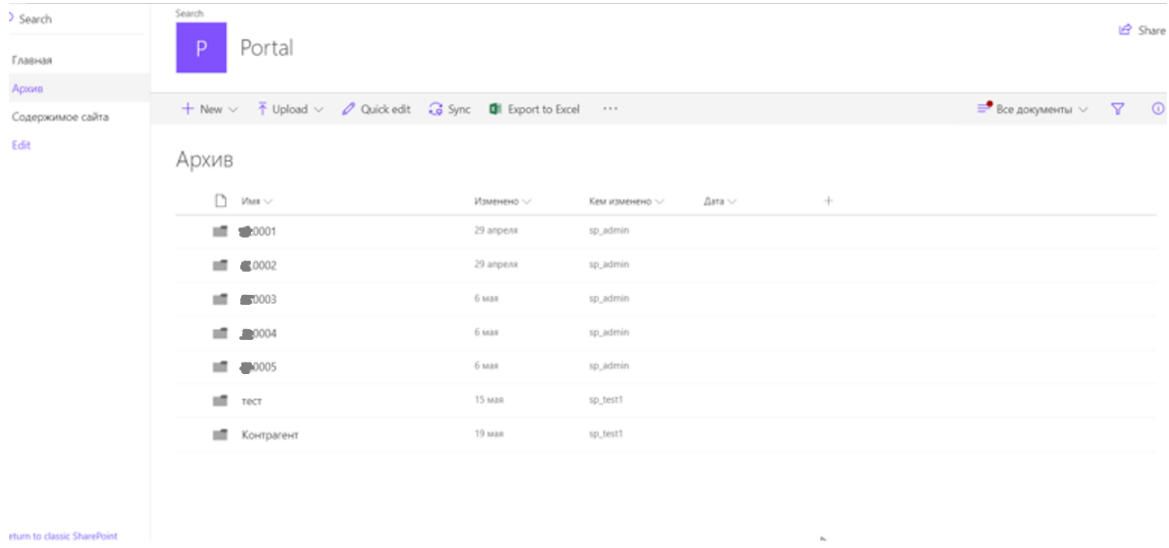
- Корпоративный портал SharePoint: возможности и функционал

Портал SharePoint не является готовым «коробочным» решением, подходящим для каждой компании. Его нужно приспосабливать под цели пользователя: выстраивать структуру, создавать отдельные узлы, страницы. В результате получится то, что требуется конкретной организации.

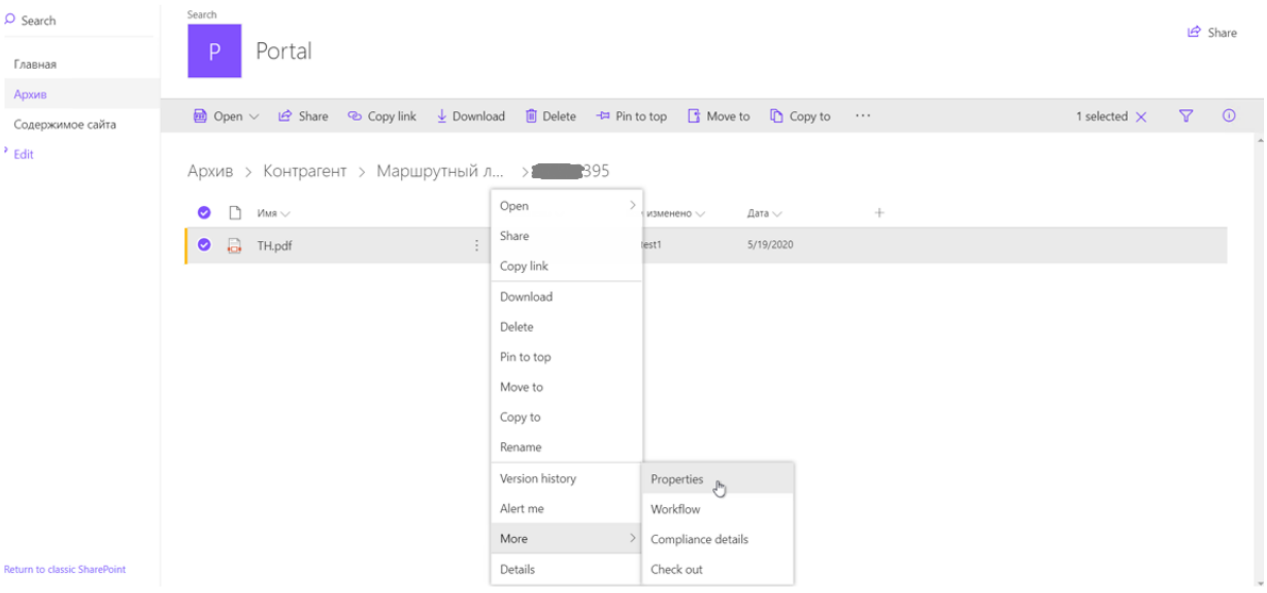
Для входа во внутрикорпоративный портал пользователю необходимо ввести свой уникальный логин и пароль в соответствующие окна. Далее необходимо нажать кнопку «Войти».

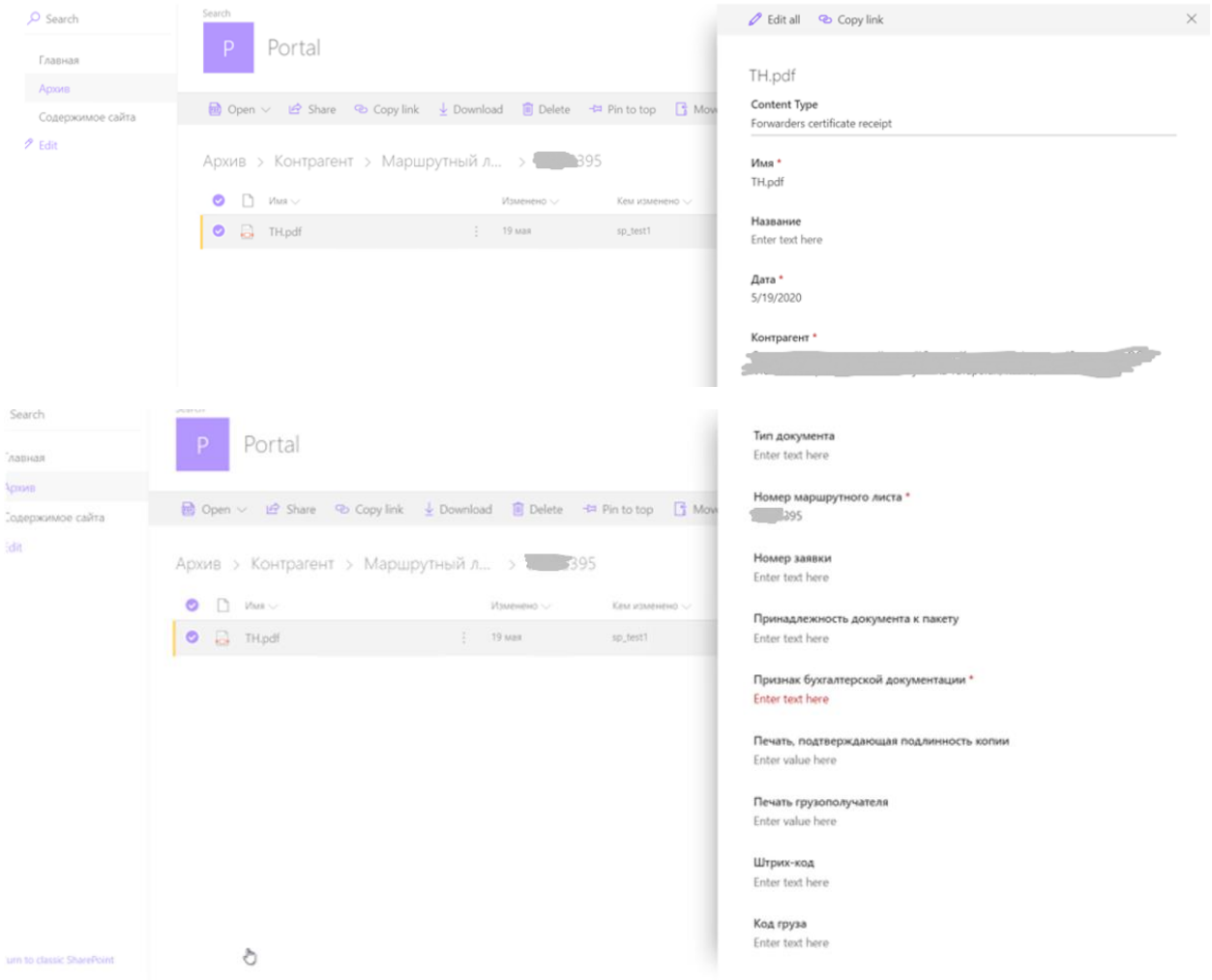


Компания самостоятельно выстраивает структуру, создает узлы и страницы под свои цели на портале.



(Рис. 125 Список каталогов в библиотеке)




Search
Главная
Архив
Содержимое сайта
Edit
Portal
Open
Share
Copy link
Download
Delete
Pin to top
Архив > Контрагент > Маршрутный л... > [redacted]395
Имя
Изменено
Кем изменено
TH.pdf
19 мая
sp_test1
Edit all
Copy link
TH.pdf
Content Type
Forwarders certificate receipt
Имя *
TH.pdf
Название
Enter text here
Дата *
5/19/2020
Контрагент *
[redacted]
Тип документа
Enter text here
Номер маршрутного листа *
[redacted]395
Номер заявки
Enter text here
Принадлежность документа к пакету
Enter text here
Признак бухгалтерской документации *
Enter text here
Печать, подтверждающая подлинность копии
Enter value here
Печать грузополучателя
Enter value here
Штрих-код
Enter text here
Код груза
Enter text here

5. Установка и удаление системы.

5.1. Системные требования

Системные требования к клиентскому компьютеру

Продукт SOICA — это набор web-приложений и сервисов, которые не требуют установки клиентской части.

Работа с системой осуществляется с помощью последней официальной версии следующих интернет-браузеров:

- Яндекс Браузер
- Атом
- Спутник
- Mozilla Firefox;
- Google Chrome;
- Apple Safari (Mac OS).

Системные требования к серверам

Ниже представлены требования к виртуальному серверу для достижения скорости обработки документов 1000-2000 страниц\час в зависимости от сложности сценария. Для увеличения производительности нужно увеличивать количество виртуальных машин.

		Требования без отказоустойчивости	**Требования с отказоустойчивостью**
Сервер:	Количество серверов	1	2
	ЦПУ	1 x Intel CPU(поддержка AVX2) – 16 ядер, 2,7 Гц и выше	1 x Intel CPU(поддержка AVX2) – 16 ядер, 2,7 Гц и выше
	ОЗУ	64ГБ	64ГБ
	НЖМД	500 ГБ (SSD)	500 ГБ (SSD)
	ОС	Windows server 2016\2019 X64 rus. Ubuntu Server 22.04 LTS Astra Linux 1.7 (при условии использования docker) РЕД ОС 7.3 (при условии использования docker) Альт Сервер 10.1 (при условии использования docker)	Windows server 2016\2019 X64 rus. Ubuntu Server 22.04 LTS Astra Linux 1.7 (при условии использования docker) РЕД ОС 7.3 (при условии использования docker) Альт Сервер 10.1 (при условии использования docker)
	Сеть	100 Мбит	100 Мбит

5.2. Установка в Linux

Поддерживаемые платформы

- Ubuntu
- Astra Linux
- РЕД ОС
- Альт Линукс

5.2.1 Установка в РЕД ОС

1. Добавляем пользователя soica в sudoers

```
# su -
# usermod -G wheel soica
```

Расскомментировать строку 'WHEEL_USERS ALL=(ALL) ALL' в файле /etc/sudoers

2.Установливаем необходимые пакеты

```
# dnf install docker-ce docker-compose
```

В конфигурационный файл демона docker добавляем следующую строку:

```
# sudo mcedit /etc/docker/daemon.json
```

```
{
   "insecure-registries": ["195.19.176.194:5000"]
}
```

3. Включаем сервис в автозагрузку системы

```
# systemctl enable docker
```

Добавить пользователя soica в группу docker:

```
# sudo groupadd docker
# sudo usermod -aG docker soica
# newgrp docker
```

4. Запуск системы Soica

Создаем каталог ./soica-ocr и извлекаем файлы из архива SoicaCore.tar и переходим в него

```
# sudo mkdir {path}
# tar -xvf SoicaCore.tar -C {path}
# cd {path}
```

При необходимости изменить порты привязки, строку авторизации в БД, и т.д. редактируем файл docker-compose.yml

```
mcedit ./docker-compose.yml
```

```
version: "3.3"
services:
  postgres:
    image: postgres:14.7
    container_name: postgres
    restart: always
    environment:
      POSTGRES_DB: "soica"
      POSTGRES_USER: "postgres"
      POSTGRES_PASSWORD: "password!1"
    volumes:
      - ./postgres/initdb/init.sql:/docker-entrypoint-initdb.d/init.sql
      - ./postgres/data:/var/lib/postgresql/data
    ports:
      - "5432:5432"
    networks:
      - default

  soica.identity:
    image: 195.19.176.194:5000/soica.identity
    container_name: soica.identity
    ports:
      - "10181:80"
```

```
    volumes:
      - ./shared_keys:/app/shared_keys
    environment:
      - ASPNETCORE_URLS=http://+:80
      - ASPNETCORE_ConnectionStrings__SoicaDB=Host=postgres;Port=5432;
Database=soica;Username=postgres;Password=password!1;Include Error Det
ail=True;
      - ASPNETCORE_AppSettings__SharedKeysFolderPath=/app/shared_keys/
    networks:
      - default

  soica.recognition:
    image: 195.19.176.194:5000/soica.recognition
    container_name: soica.recognition
    restart: always
    ports:
      - "10184:80"
    volumes:
      - ./normalization:/app/normalization
      - ./tmp:/app/tmp
      - ./logs:/app/logs
    environment:
      - ASPNETCORE_URLS=http://+:80
    networks:
      - default

  soica.web-api:
    image: 195.19.176.194:5000/soica.web-api
    container_name: soica.web-api
    ports:
      - "10188:80"
    volumes:
      - ./shared_keys:/app/shared_keys
      - ./file_storage:/app/file_storage
```

```
      - ./shared_import:/app/shared_import
      - ./shared_export:/app/shared_export
      - ./shared_data_sources:/app/shared_data_sources
      - ./ya_vision:/app/ya_vision
      - ./logs:/app/logs
      - ./tesseract:/app/tesseract
    environment:
      - ASPNETCORE_URLS=http://+:80
      - ASPNETCORE_ConnectionStrings__SoicaDB=Host=postgres;Port=5432;
Database=soica;Username=postgres;Password=password!1;Include Error Det
ail=True;
      - ASPNETCORE_AppSettings__soica_web_uri=http://soica.recognition
/
    extra_hosts:
      - localhost:host-gateway
    networks:
      - default


  soica.control-panel:
    image: 195.19.176.194:5000/soica.control-panel
    container_name: soica.control-panel
    ports:
      - "10182:80"
    volumes:
      - ./shared_keys:/app/shared_keys
      - ./file_storage:/app/file_storage
      - ./shared_import:/app/shared_import
      - ./shared_export:/app/shared_export
      - ./shared_data_sources:/app/shared_data_sources
      - ./ya_vision:/app/ya_vision
      - ./logs:/app/logs
      - ./tesseract:/app/tesseract
    environment:
      - ASPNETCORE_URLS=http://+:80
```

```
      - ASPNETCORE_ConnectionStrings__SoicaDB=Host=postgres;Port=5432;
Database=soica;Username=postgres;Password=password!1;Include Error Det
ail=True;
      - ASPNETCORE_AppSettings__soica_web_uri=http://soica.recognition
/
      - ASPNETCORE_AppSettings__service_address=http://soica.web-api/
      - ASPNETCORE_AppSettings__validation_url=http://localhost/
      - ASPNETCORE_AppSettings__datasource_path=/app/shared_data_sourc
es
    extra_hosts:
      - localhost:host-gateway
    networks:
      - default


  soica.validation:
    image: 195.19.176.194:5000/soica.validation
    container_name: soica.validation
    environment:
      - ASPNETCORE_URLS1=https://+:443;http://+:80
      - ASPNETCORE_ConnectionStrings__SoicaDB=Host=postgres;Port=5432;
Database=soica;Username=postgres;Password=password!1;Include Error Det
ail=True;
      - ASPNETCORE_IdentityService__Uri=http://identyty_external_host/
    ports:
      - "10183:80"
      - "10143:443"
    volumes:
      - ./shared_keys:/app/shared_keys
      - ./file_storage:/app/file_storage
      - ./logs:/app/logs
    extra_hosts:
      - "localhost:host-gateway"
    networks:
      - default
```

Авторизируемся в реестре Docker Soica

```
# docker login http://195.19.176.194:5000/
```

Вводим имя пользователя и пароль

Запускаем сервис postgres в контейнере

```
# docker-compose up -d postgres
```

Запускаем сервисы сойка

```
# docker-compose up -d
```

Вариант работы системы через балансировщик нагрузки

```
version: "3.3"
services:
  soica.identity:
    image: 195.19.176.194:5000/soica.identity
    container_name: soica.identity
    ports:
      - "10181:80"
    volumes:
      - ./shared_keys:/app/shared_keys
    environment:
      - ASPNETCORE_URLS=http://+:80
      - ASPNETCORE_ConnectionStrings__SoicaDB=Host=postgres;Port=5432;
Database=soica;Username=postgres;Password=password!1;Include Error Det
ail=True;
      - ASPNETCORE_AppSettings__SharedKeysFolderPath=/app/shared_keys/
    networks:
      - default


  soica.recognition.1:
```

```
    image: 195.19.176.194:5000/soica.recognition
    container_name: soica.recognition.1
    restart: always
    ports:
      - "3001:80"
    volumes:
      - ./normalization:/app/normalization
      - ./tmp:/app/tmp
      - ./logs:/app/logs
    environment:
      - ASPNETCORE_URLS=http://+:80
    networks:
      - default

  soica.recognition.2:
    image: 195.19.176.194:5000/soica.recognition
    container_name: soica.recognition.2
    restart: always
    ports:
      - "3002:80"
    volumes:
      - ./normalization:/app/normalization
      - ./tmp:/app/tmp
      - ./logs:/app/logs
    environment:
      - ASPNETCORE_URLS=http://+:80
    networks:
      - default

  soica.recognition.2:
    image: 195.19.176.194:5000/soica.recognition
    container_name: soica.recognition.2
    restart: always
```

```
    ports:
      - "3002:80"
    volumes:
      - ./normalization:/app/normalization
      - ./tmp:/app/tmp
      - ./logs:/app/logs
    environment:
      - ASPNETCORE_URLS=http://+:80
    networks:
      - default


  soica.web-api:
    image: 195.19.176.194:5000/soica.web-api
    container_name: soica.web-api
    ports:
      - "10188:80"
    volumes:
      - ./shared_keys:/app/shared_keys
      - ./file_storage:/app/file_storage
      - ./shared_import:/app/shared_import
      - ./shared_export:/app/shared_export
      - ./shared_data_sources:/app/shared_data_sources
      - ./ya_vision:/app/ya_vision
      - ./logs:/app/logs
      - ./tesseract:/app/tesseract
    environment:
      - ASPNETCORE_URLS=http://+:80
      - ASPNETCORE_ConnectionStrings__SoicaDB=Host=postgres;Port=5432;
Database=soica;Username=postgres;Password=password!1;Include Error Det
ail=True;
      - ASPNETCORE_AppSettings__soica_web_uri=http://soica.recognition
/
    extra_hosts:
      - localhost:host-gateway
```

```
    networks:
      - default


  soica.control-panel:
    image: 195.19.176.194:5000/soica.control-panel
    container_name: soica.control-panel
    ports:
      - "10182:80"
    volumes:
      - ./shared_keys:/app/shared_keys
      - ./file_storage:/app/file_storage
      - ./shared_import:/app/shared_import
      - ./shared_export:/app/shared_export
      - ./shared_data_sources:/app/shared_data_sources
      - ./ya_vision:/app/ya_vision
      - ./logs:/app/logs
      - ./tesseract:/app/tesseract
    environment:
      - ASPNETCORE_URLS=http://+:80
      - ASPNETCORE_ConnectionStrings__SoicaDB=Host=postgres;Port=5432;
Database=soica;Username=postgres;Password=password!1;Include Error Det
ail=True;
      - ASPNETCORE_AppSettings__soica_web_uri=http://soica.recognition
/
      - ASPNETCORE_AppSettings__service_address=http://soica.web-api/
      - ASPNETCORE_AppSettings__validation_url=http://localhost/
      - ASPNETCORE_AppSettings__datasource_path=/app/shared_data_sourc
es
    extra_hosts:
      - localhost:host-gateway
    networks:
      - default


  soica.validation:
```

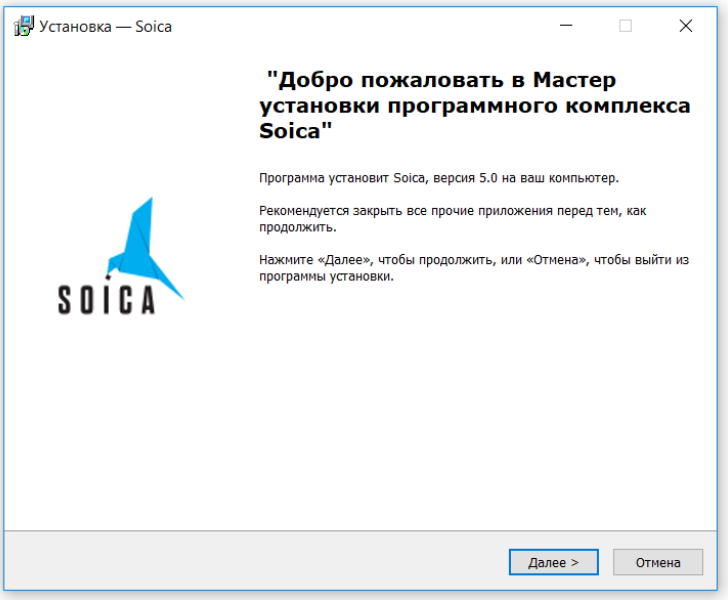
```
    image: 195.19.176.194:5000/soica.validation
    container_name: soica.validation
    environment:
      - ASPNETCORE_URLS1=https://+:443;http://+:80
      - ASPNETCORE_ConnectionStrings__SoicaDB=Host=postgres;Port=5432;
Database=soica;Username=postgres;Password=password!1;Include Error Det
ail=True;
      - ASPNETCORE_IdentityService__Uri=http://identyty_external_host/
    ports:
      - "10183:80"
      - "10143:443"
    volumes:
      - ./shared_keys:/app/shared_keys
      - ./file_storage:/app/file_storage
      - ./logs:/app/logs
    extra_hosts:
      - "localhost:host-gateway"
    networks:
      - default


load-balancer:
  image: nginx:latest
  container_name: load-balancer
  volumes:
    - ./nginx/nginx.conf:/etc/nginx/nginx.conf
  ports:
    - "10184:8080"
  extra_hosts:
    - "localhost:host-gateway"
  networks:
    - default
```

5.3. Установка в Windows

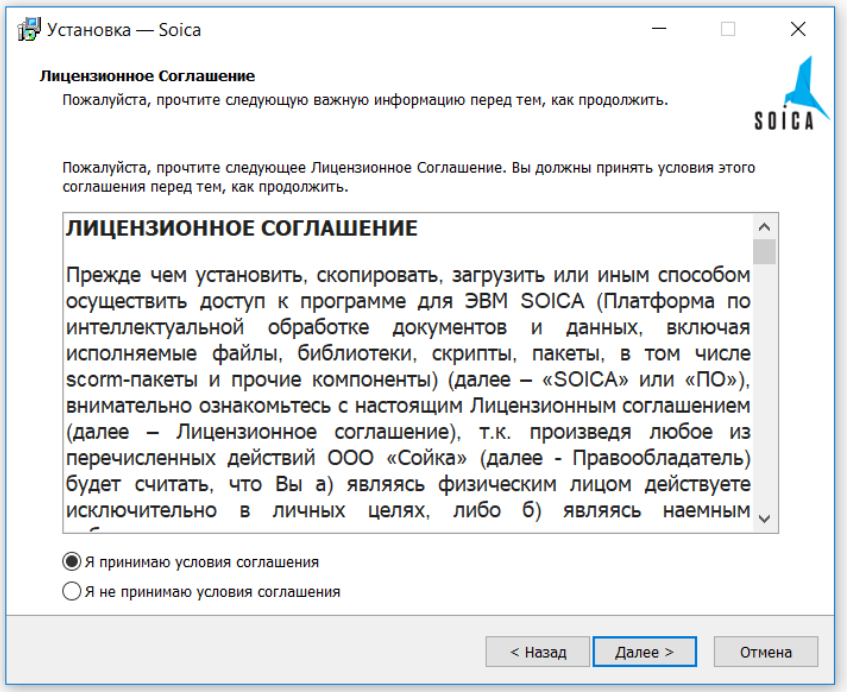
1. Запустить инсталлятор SoicaService.exe
2. Для работы системы Soica должны быть установлены следующие программы/компоненты:

a. PostgreSQL v.11
b. IIS
c. .Net Framework 4.6.1
d. Mongo DB

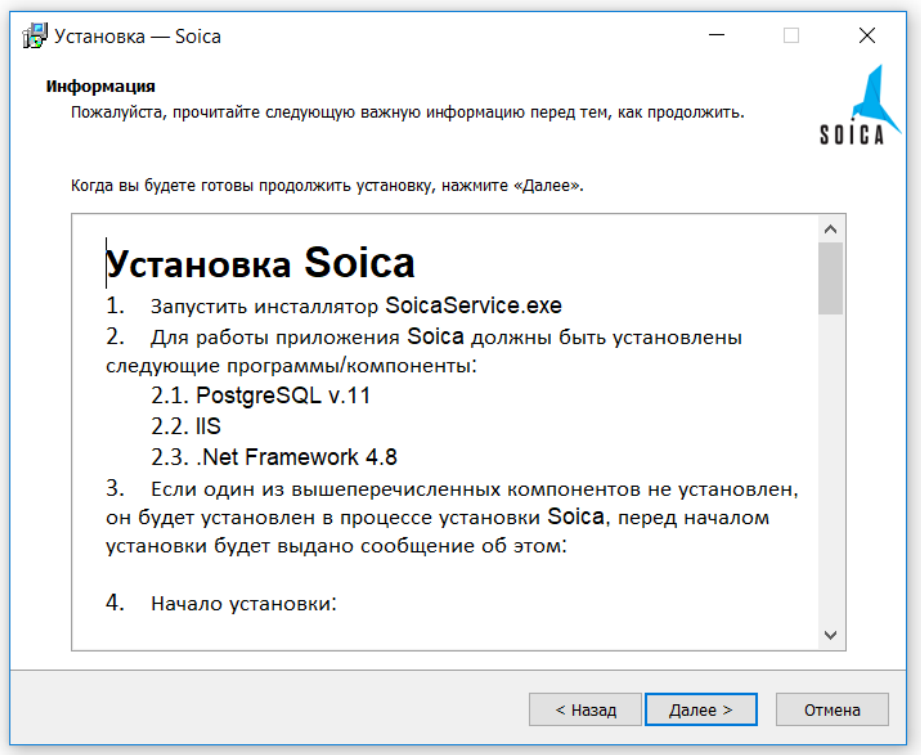
3. Начало установки:



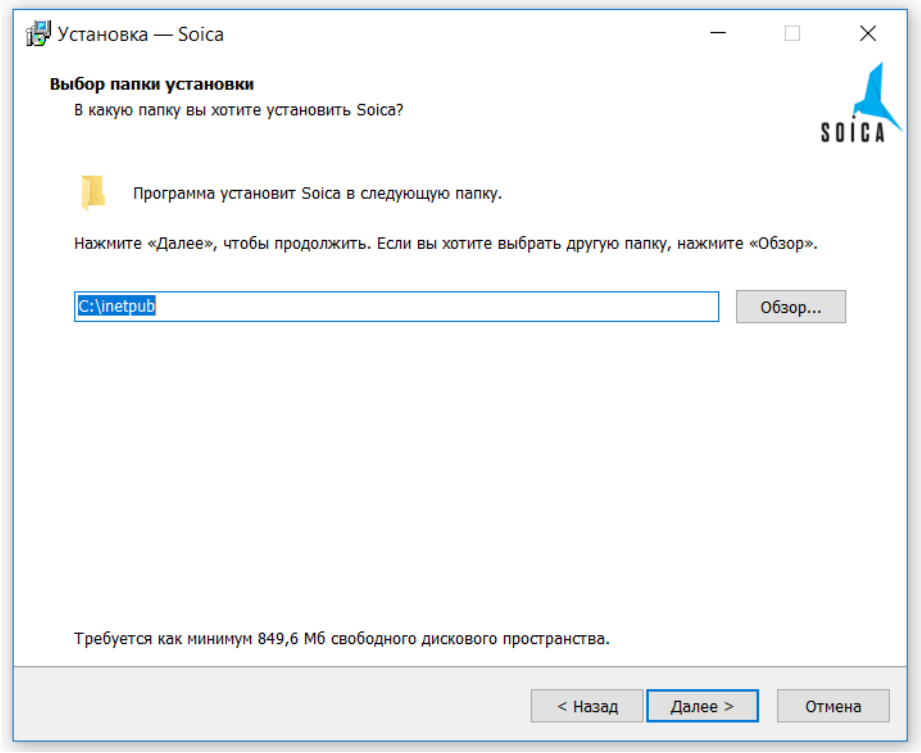
4. Необходимо принять условия соглашения и нажать кнопку «Далее»



5. Если одного или несколько из перечисленных в п. 2 компонентов нет на локальном компьютере, они будет установлены автоматически. Перед началом установки системы Soica будет выдано сообщение об этом:

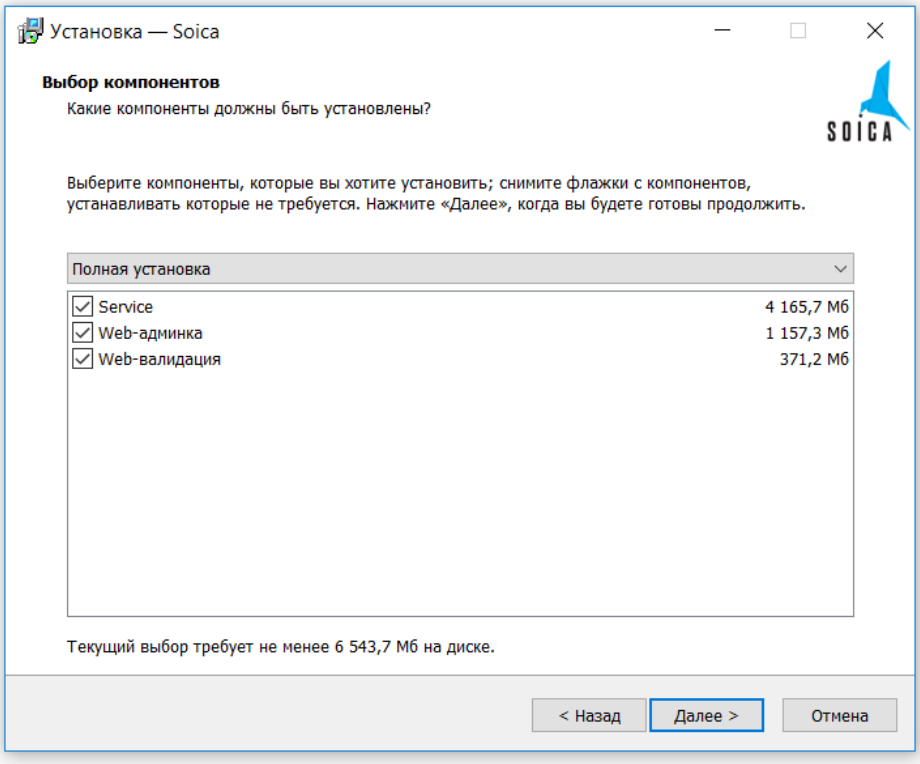


6. Необходимо указать папку, в которой будет развернут сервис Soica:



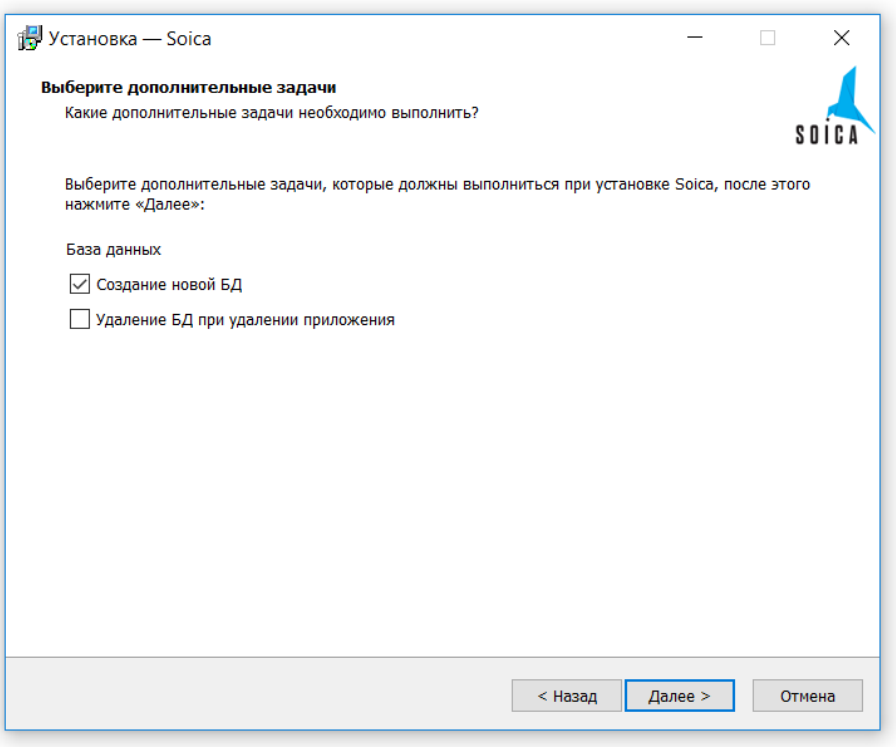
Папки для установки Web-приложений: администратор - C:\inetpub\AdministratorWebUI и валидация - C:\inetpub\Validation заданы по умолчанию.

7. Укажите нужные компоненты:

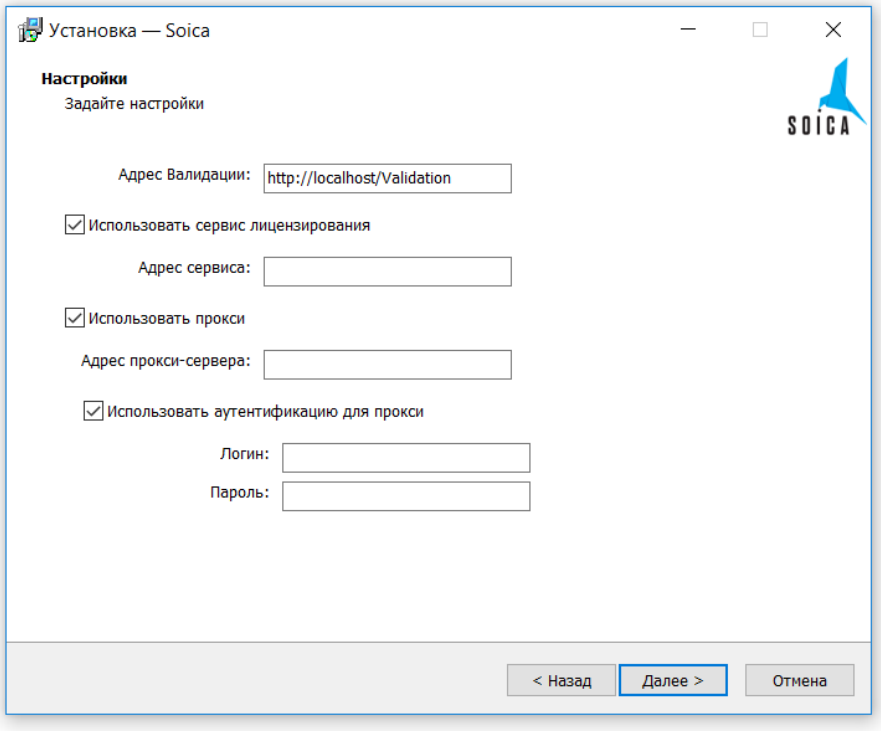


8. Выбрать дополнительные задачи.

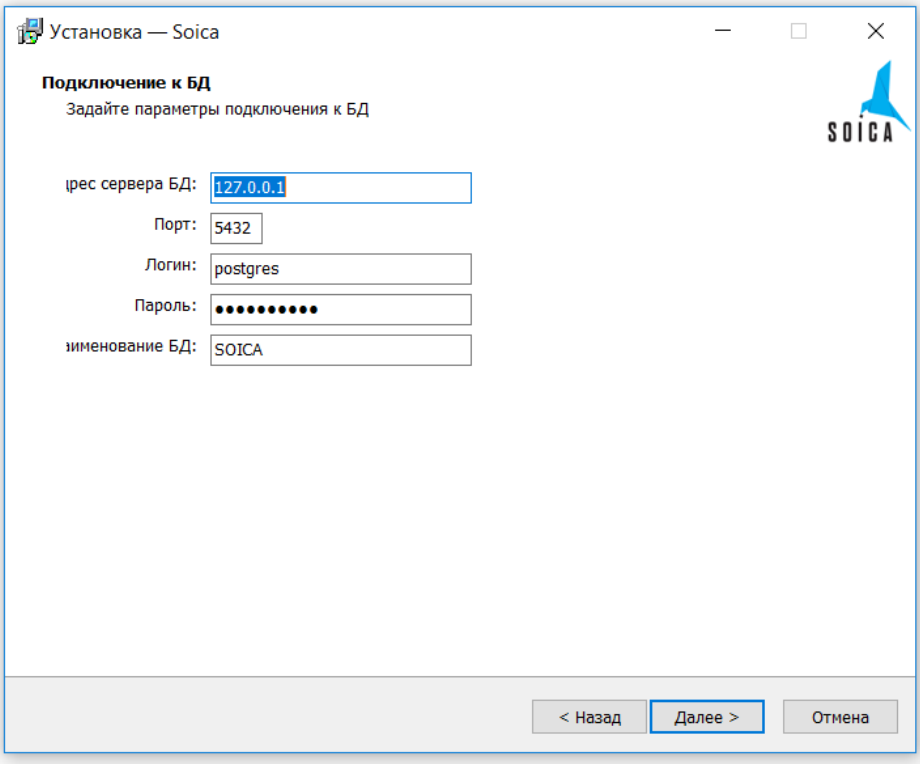
Если PostgreSQL не установлен на вашем локальном компьютере, Вам будет предложено установить его. Также можно развернуть БД (Создание БД). **Не выбирайте этот пункт в том случае, если Вам нужно подключится к другому серверу БД.** Параметры подключения Вы сможете указать на следующем шаге установки.



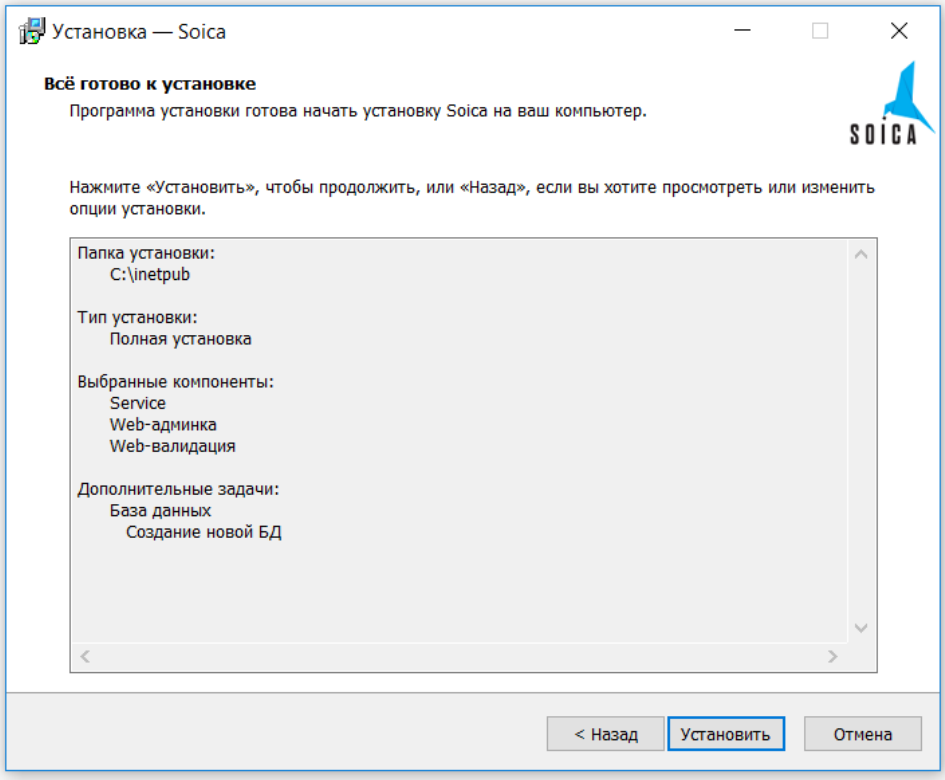
9. Задайте настройки



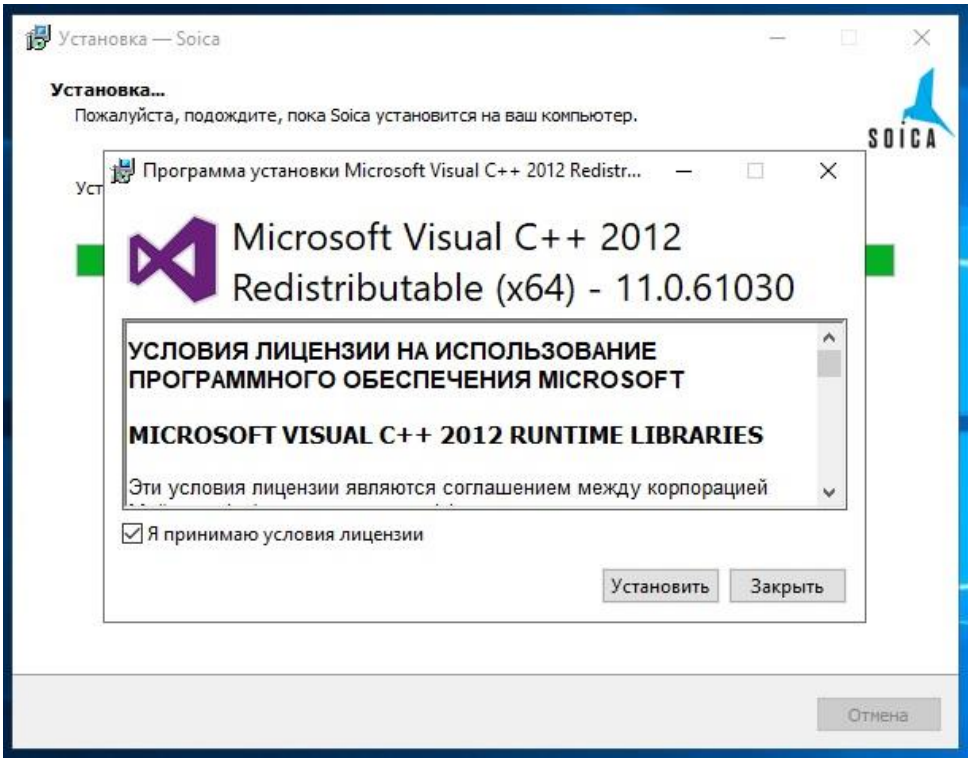
10. Задать параметры подключения к БД:



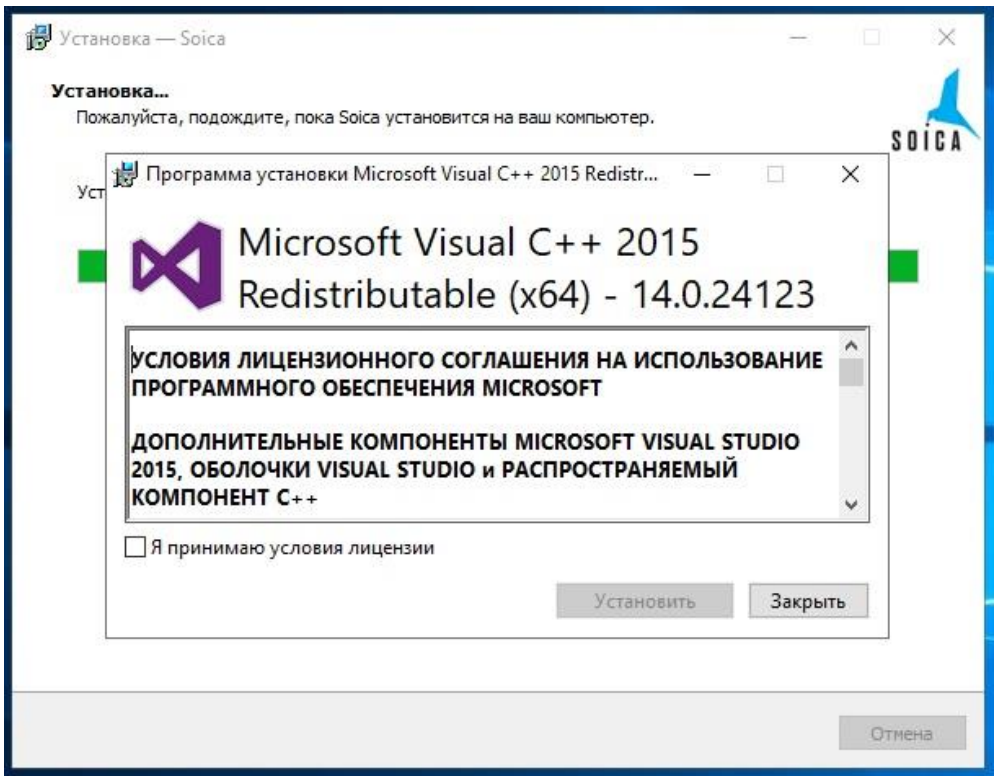
11. Страница выбранных компонентов. На представленной форме выведены все выбранные настройки и компоненты для установки:



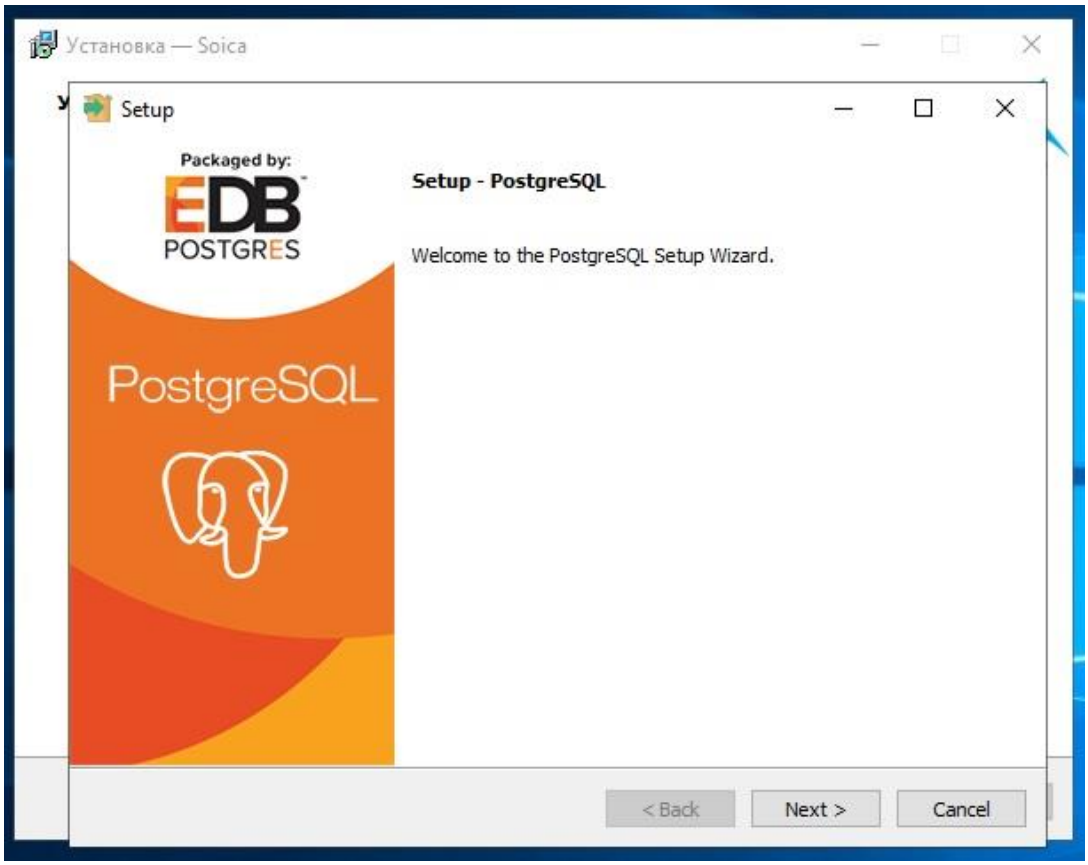
12. После основной установки происходит запуск установки необходимых программ/компонентов:
Распространяемый пакет Visual C++ для Visual Studio 2012 обновление 4



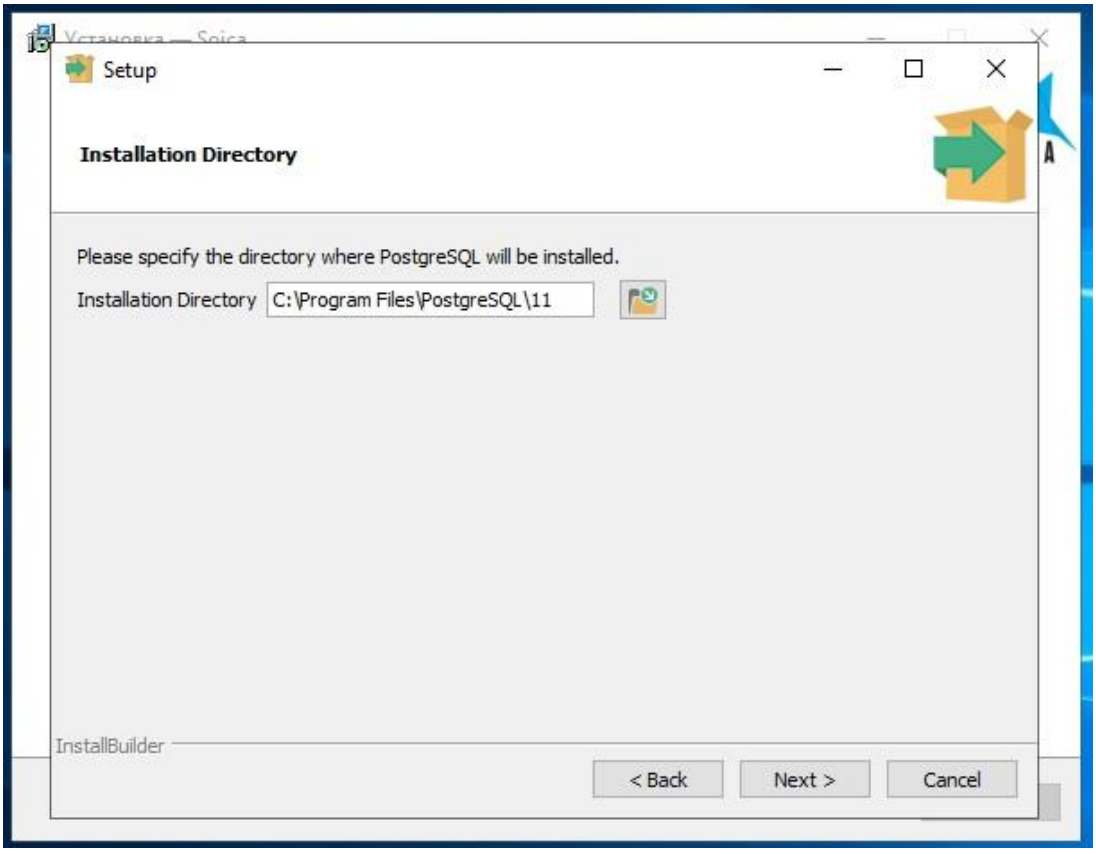
Распространяемый компонент Microsoft Visual C++ 2015 Update 3 RC



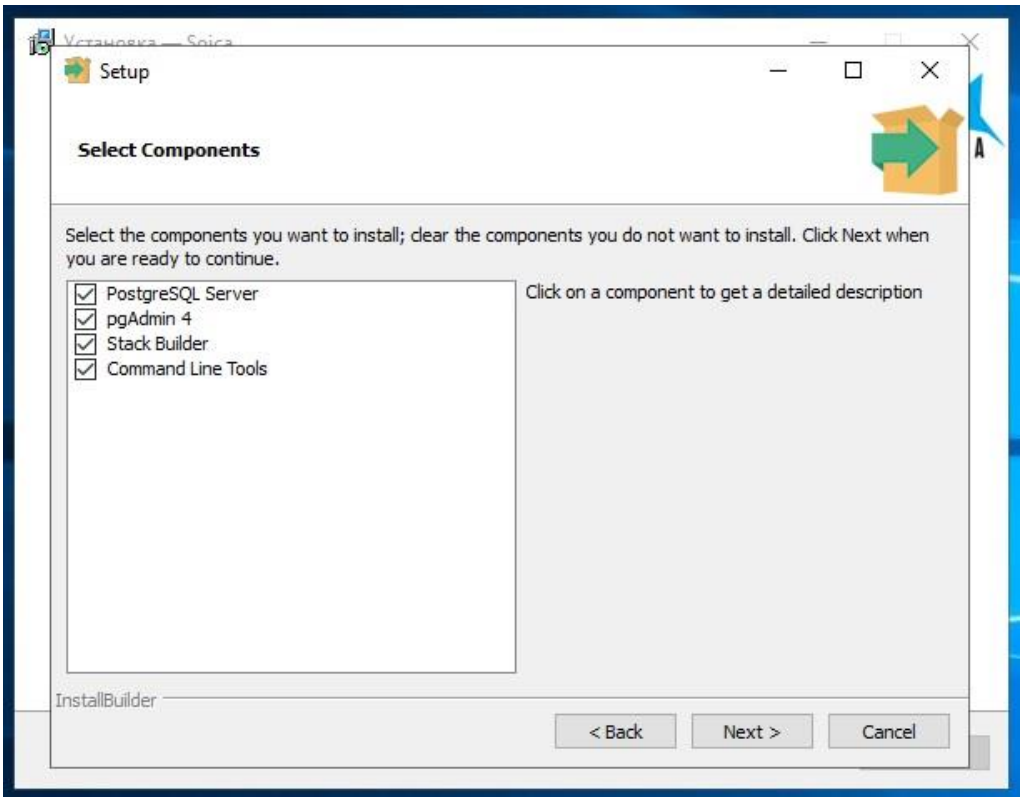
Установка PostgreSQL v.11 в случае, если PostgreSQL не установлен:



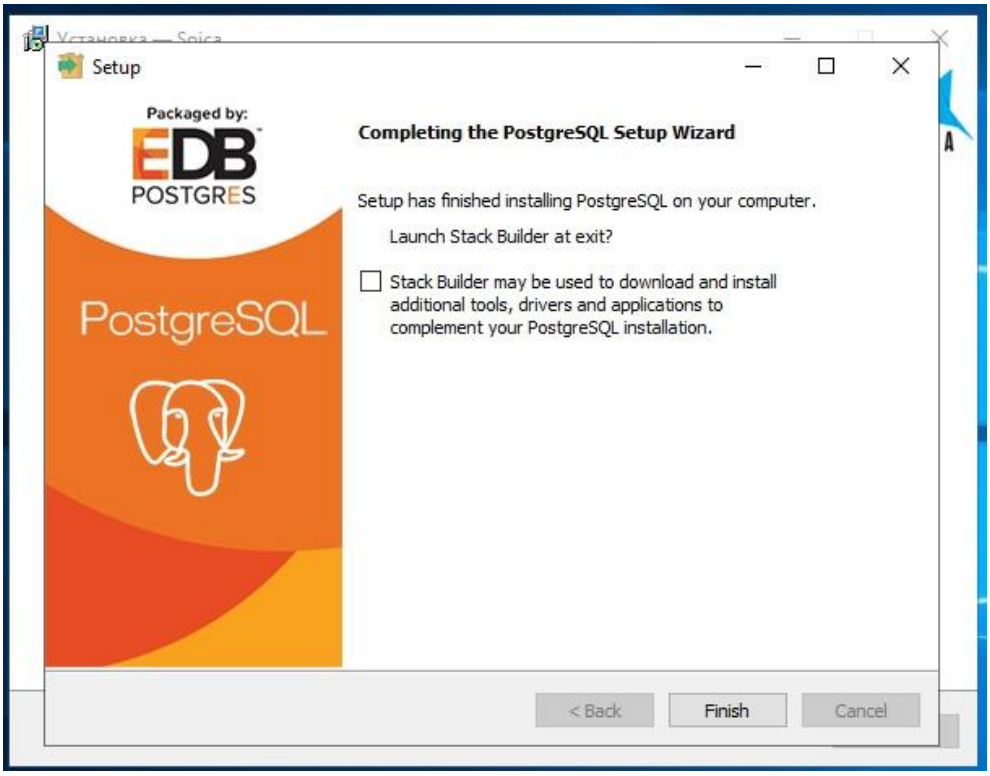
Выберите папку для установки PostgreSQL:



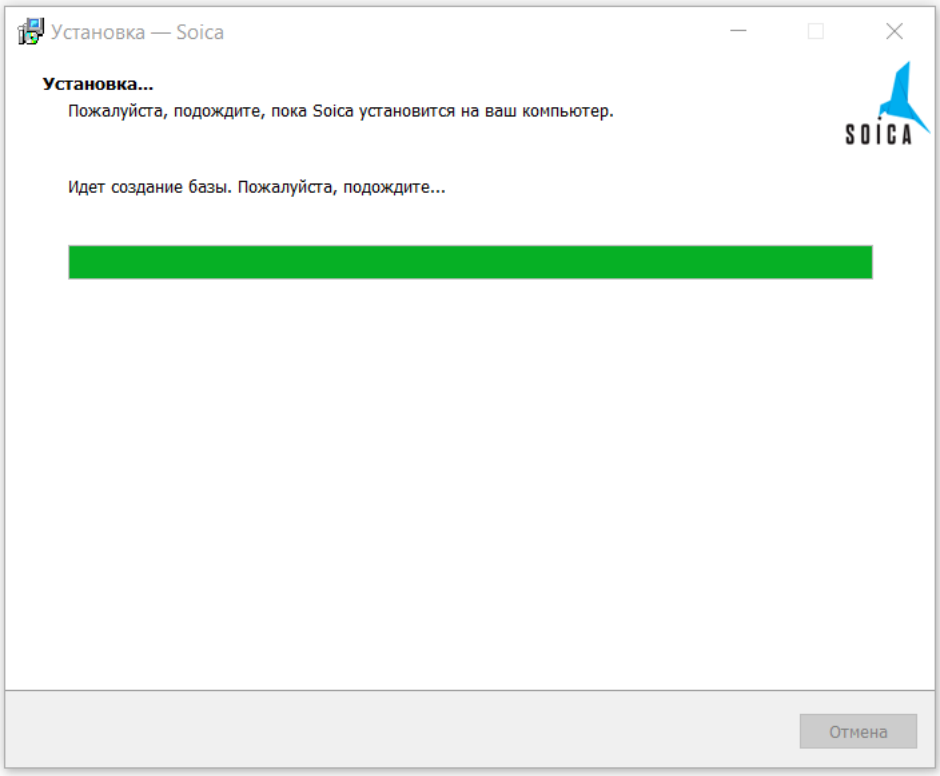
Выберите все компоненты для установки:



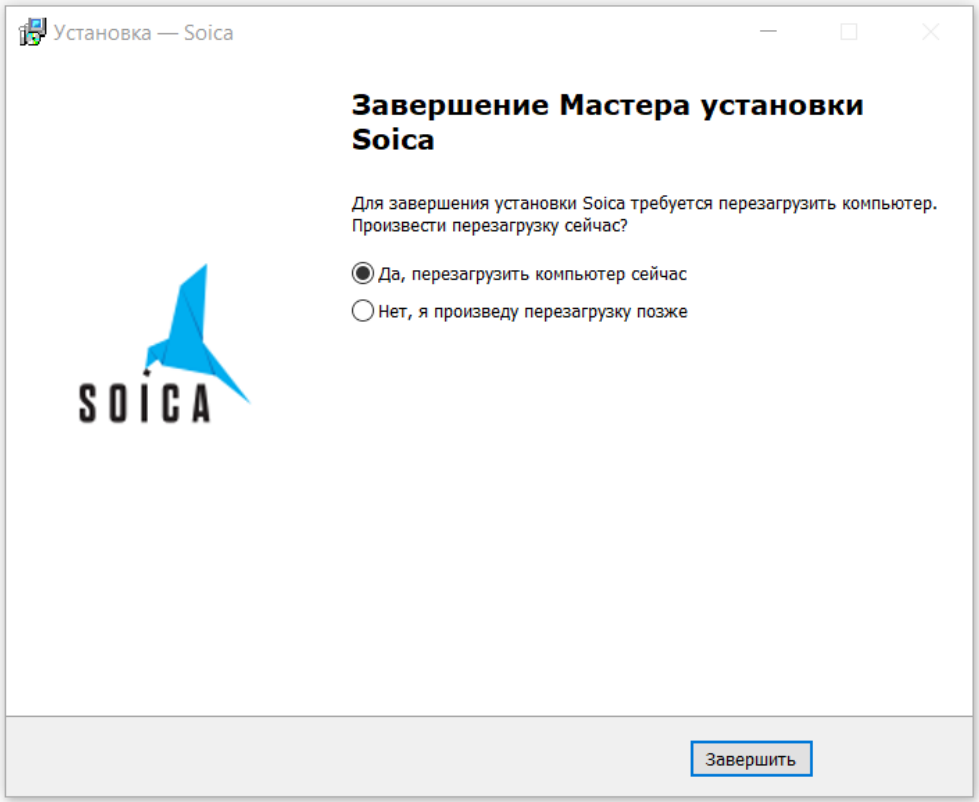
После завершения установки снимите галочку для того, чтобы не запускать StackBuilder после установки



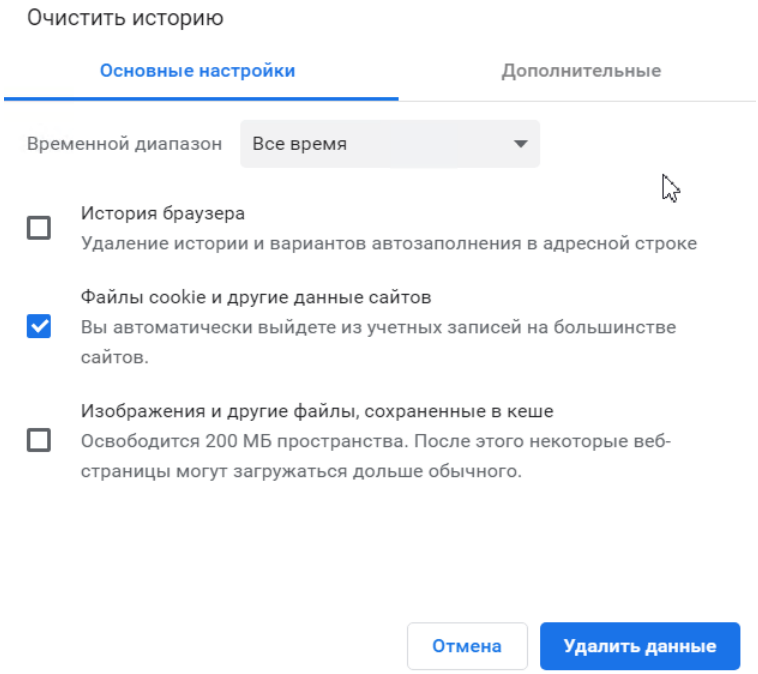
13. После установки всех дополнительных компонент происходит создание базы и настройка web-приложений в IIS:



14. Завершение установки. После удачной установки появится форма завершения установки и запрос перезагрузки компьютера. При нажатии на кнопку «Завершить» инсталлятор закончит свою работу и перезагрузит компьютер:



15. После перезагрузки необходимо очистить куки браузера. Открыть браузер, зайти в настройки и очистить историю.



***Форма может отличатся в зависимости от браузера.**

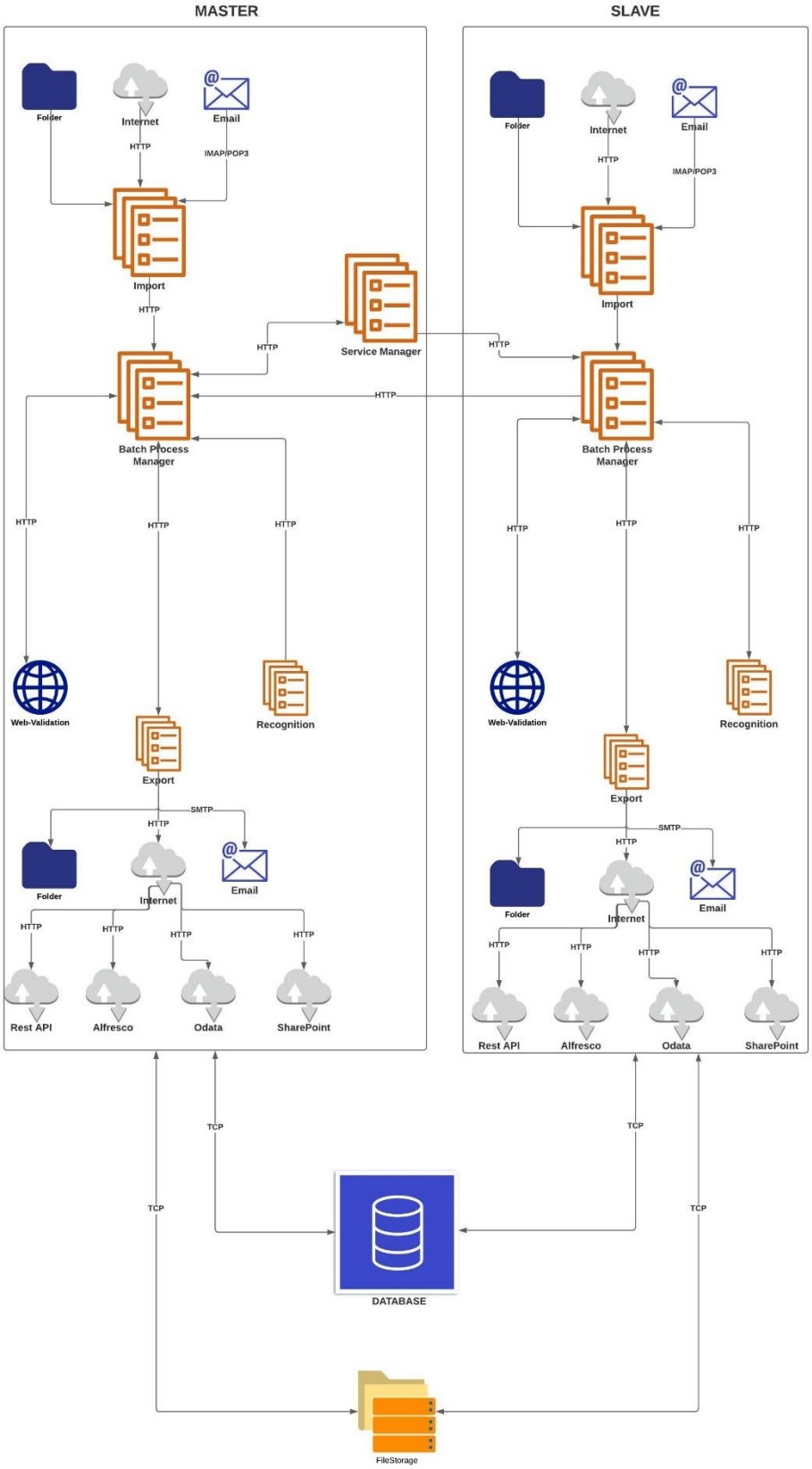
16. После установки модуль администратора будет доступен на локальном компьютере по ссылке: http://localhost/administrator Модуль валидации по ссылке http://localhost/Validation

5.4. Распределенная установка на ОС Windows

Распределенная установка – это установка одной системы на нескольких серверах. При этом один сервер является ведущим (master), остальные являются подчиненными (slave).

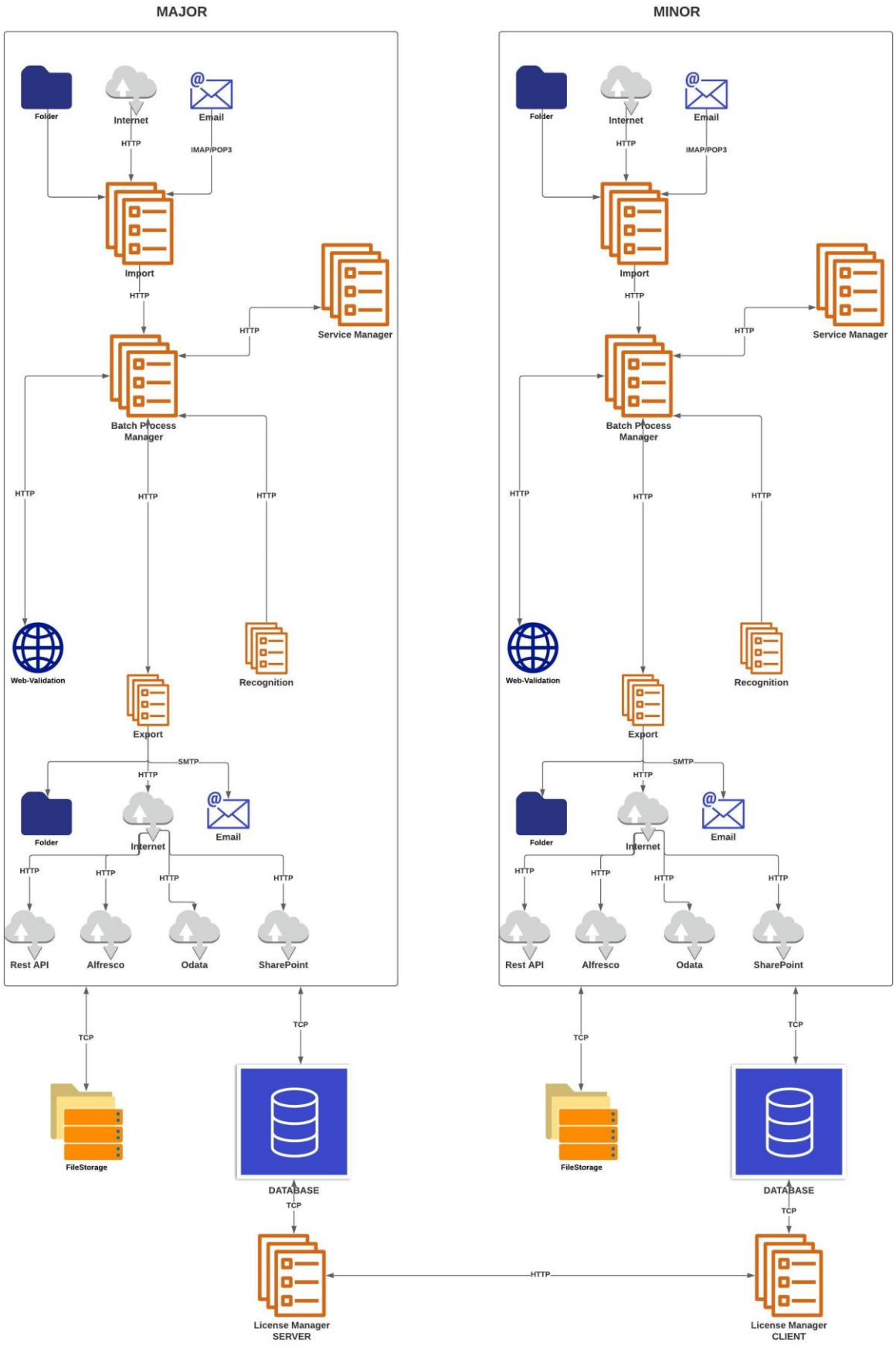
Ведущий сервер распределяет нагрузку между всеми серверами, участвующими в обработке пакетов. Пакет отправляется на обработку наименее загруженному (загрузка процессора) серверу. Таким образом несколько пакетов могут одновременно обрабатываться на нескольких серверах.

Архитектура распределенной установки:



Сервисы с единым пулом лицензий

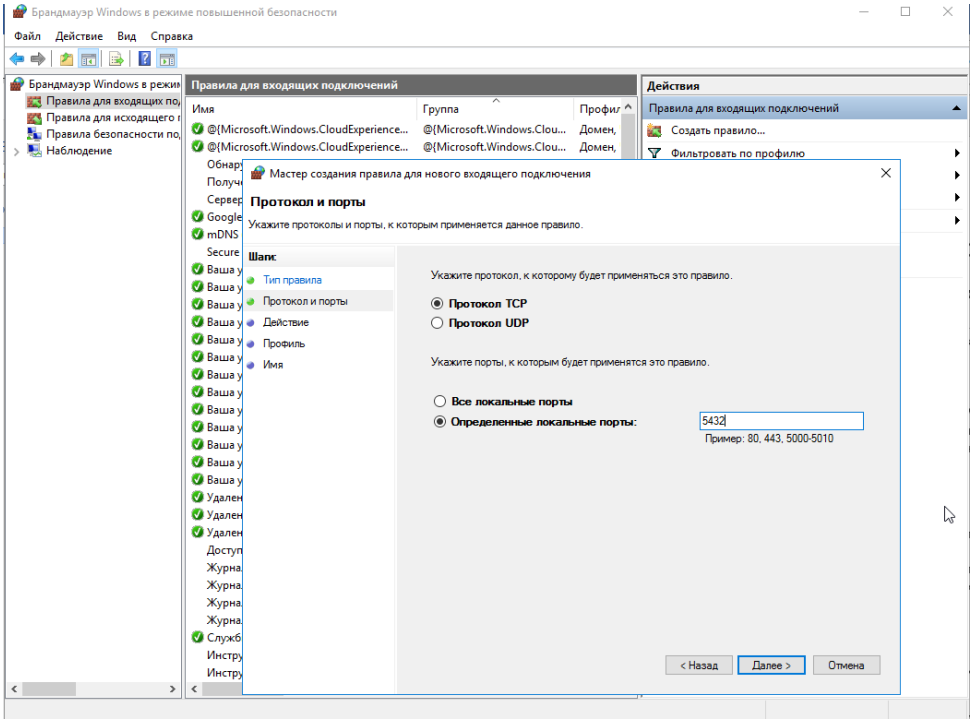
Архитектура независимой работы сервисов с единым пулом лицензий:



Один из вариантов распределенной установки – установка системы на нескольких независимых серверах, но с использованием одного пула приложений. При этом один сервер выполняет роль основного (MAJOR), а остальные сервера – второстепенные (MINOR). Второстепенные сервера синхронизируют лицензии с основным сервером при помощи сервиса лицензирования. Периодичность синхронизации настраивается в конфигурации второстепенных сервисов.

Настройка удаленного доступа к серверу Postgres:

1. На сервере Postgres открыть порт 5432, для этого в Брандмауэр Windows добавить его в Правила входящих подключений/Правила исходящих подключений.



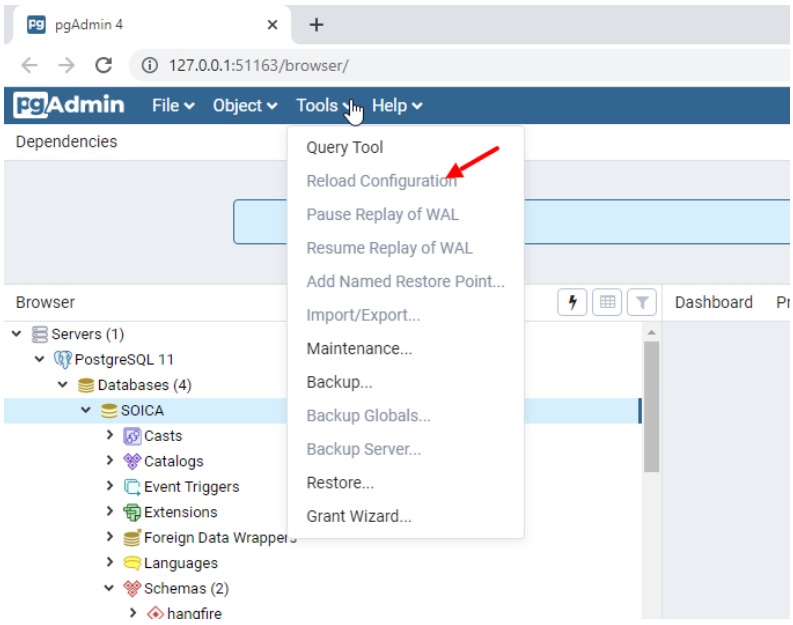
2. Открыть файл **pg_hba. conf** (находится в папке расположения PG) и добавить туда строку:

```
# IPv4 local connections:
host    all             all             127.0.0.1/32            md5
host    all             all             192.168.0.0/0           md5
```

```
#
# This file is read on server startup and when the server receives a
# SIGHUP signal.  If you edit the file on a running system, you have to
# SIGHUP the server for the changes to take effect, run "pg_ctl reload",
# or execute "SELECT pg_reload_conf()".
#
# Put your actual configuration here
# ----------------------------------
#
# If you want to allow non-local connections, you need to add more
# "host" records.  In that case you will also need to make PostgreSQL
# listen on a non-local interface via the listen_addresses
# configuration parameter, or via the -i or -h command line switches.



# TYPE  DATABASE        USER            ADDRESS                 METHOD

# IPv4 local connections:
host    all             all             127.0.0.1/32            md5
host    all             all             192.168.0.0/0            md5
# IPv6 local connections:
host    all             all             ::1/128                 md5
# Allow replication connections from localhost, by a user with the
# replication privilege.
host    replication     all             127.0.0.1/32            md5
host    replication     all             ::1/128                 md5

```

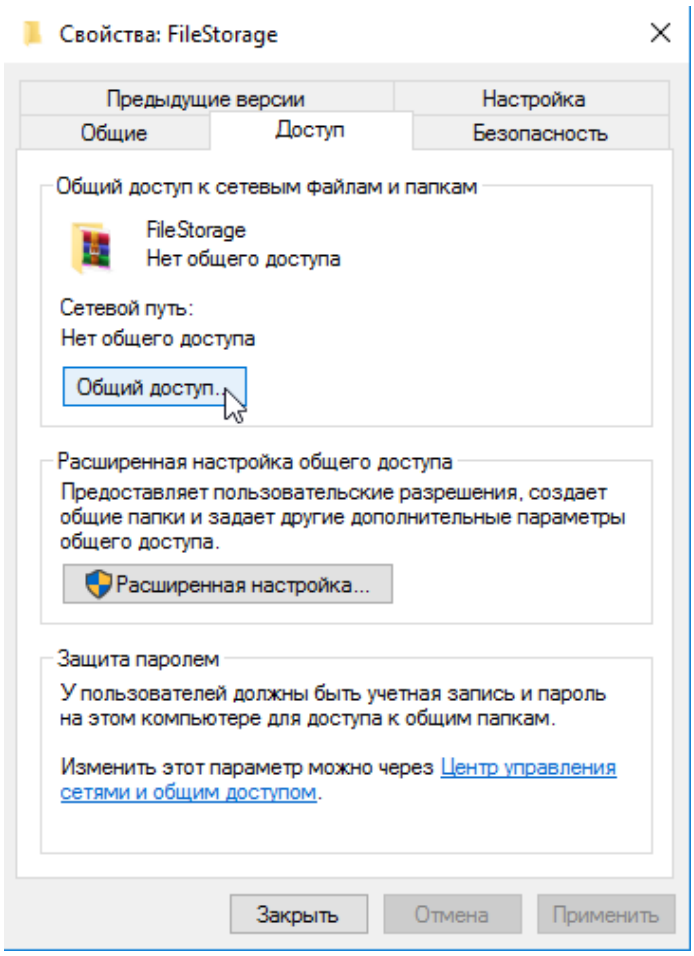
Данная строка дает право на подключение с серверов, начинающихся на адрес: 192.168.__.__

3. После этого в PG в разделе «Tools» нажать «Reload Configuration». для вступления изменений в силу

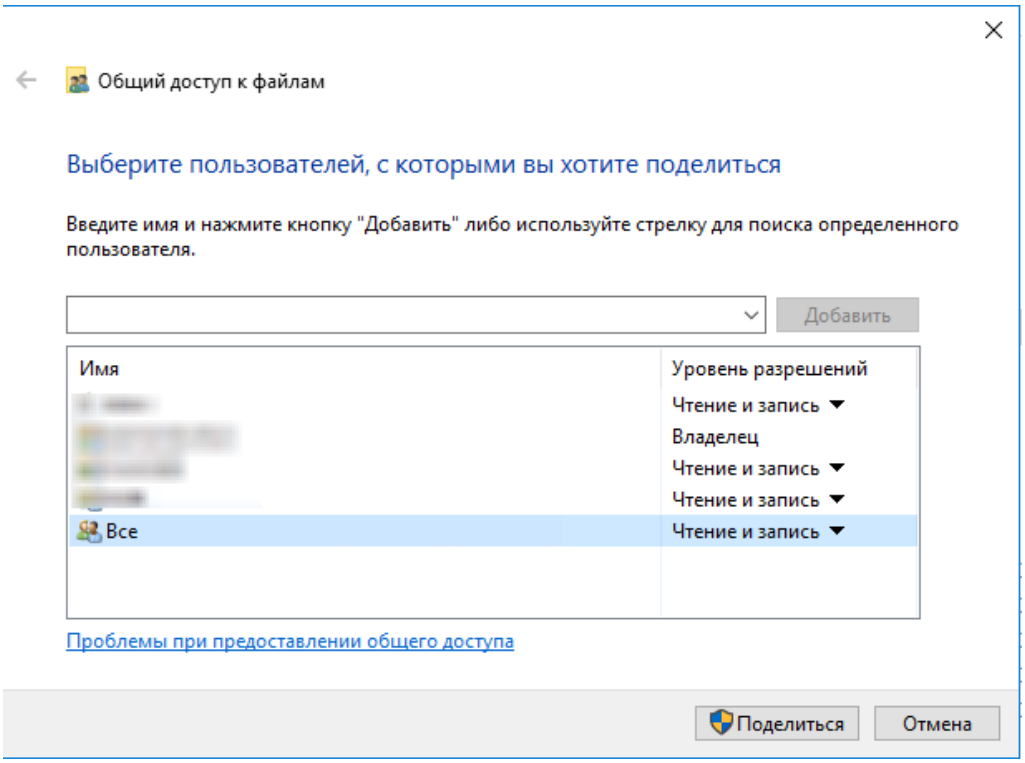


Настройка сетевых папок:

1. На одном из серверов создать папку для FileStorage и настроить сетевой доступ к этой папке.



2. На одном из серверов создать папку для размещения файла лицензии, поместить туда файл с лицензией (лицензия должна соответствовать лицензии в БД) и настроить сетевой доступ к этой папке.



Настройка конфигурации приложений:

В папках «Administrator», «SoicaWebService», «Validation» на всех серверах в **web.config** сделать следующее:

1. изменить строку подключения к БД (указать подключение к сетевой БД);
2. в параметре **lic_path** указать путь к сетевой папке с лицензией;
3. в параметре **file_storage_path** указать путь к сетевой папке FileStorage.

В папке “SoicaWebService” slave-серверов в **web.config** нужно указать адрес master-сервера в параметре **server**

```
</configSections>
<appSettings>
  <add key="DB_path" value="database=test;server=192.168.[redacted];port=5432;User ID=postgres;password=[redacted];Timeout=300;CommandTimeout=300;" />
  <add key="gs_path_64" value="C:\inetpub\SoicaWebService\bin\gsdll64.dll" />
  <add key="gs_path_32" value="C:\inetpub\SoicaWebService\bin\gsdll32.dll" />
  <add key="lic_path" value="\\[redacted]TEST-SOICA1\FileStorage[redacted]"/><!--"C:\inetpub\SoicaWebService" -->
  <add key="aspnet:UseTaskFriendlySynchronizationContext" value="true" />
  <add key="server" value="http://192.168.[redacted]/soicaservice/" />
  <add key="service_address" value="http://localhost/soicaservice/" />
  <add key="webserver" value="http://localhost/soicaservice/" />
  <add key="DBaseMode" value="single" />
  <add key="ClientSettingsProvider.ServiceUri" value="" />
  <add key="yaVisionSettingsPath" value="C:\inetpub\SoicaWebService\bin\tessdata" />
  <add key="tess5Path" value="C:\Program Files\Tesseract-OCR\" />
  <add key="soicaII" value="C:\inetpub\SoicaWebService\bin\x64" />
  <add key="rec_temp_path" value="C:\inetpub\SoicaWebService" />
  <add key="file_storage_path" value="\\[redacted]TEST-SOICA1\FileStorage207"/><!--"C:\inetpub\FileStorage"/> -->
  <!--<add key="soicaII_dict" value="C:\inetpub\SoicaWebService\bin\x64" />-->
  <add key="soicaII_maxCores" value="8" />
  <add key="soicaII_coresToBeUsed" value="8" />
  <add key="prometheus_exporter" value="localhost:1234" />
  <!--<add key="license_server" value="http://192.168.[redacted]/soicaservice/"/>-->
</appSettings>
<connectionStrings>
  <add name="TempDataModel" connectionString="Host=192.168.[redacted];Database=test;Port=5432;Username=postgres;Password=[redacted]" providerName="Npgsql" />
</connectionStrings>
```

Настройка БД:

В БД в таблицу «**services**» необходимо добавить модули с адресами серверов в системе, которые участвуют в процессе обработки документов.

Пример запроса для добавления сервисов:

INSERT INTO public.services(module_id, address) VALUES (2, ‘http://192.168.0.0/soicaservice’);

Идентификаторы модулей module_id:

- 1 - импорт;
- 2 - распознавание;
- 3 - валидация;
- 4 - экспорт.

5.5. Удаление

Удаление системы Soica и всех ее модулей производится через «Установку и удаление программ». Выберите Soica v1.0 и нажмите “Удалить”. С вашего компьютера будет удален сервис Soica, база данных SOICA, удалены из IIS web-приложения administrator и validation.