Исида-Информатика

OCR Gate for ISIDA Retriever
Установка и настройка

Код документа: 3363-2.5.15. Версия документа: 3
Дата редакции документа: 01.09.2020. Количество листов: 9

Витебск, 2020 г.

Содержание

© ООО «Исида-Информатика»			
Код	*Наименование*	*Лист*	*Листов*
3363-2.5.15	OCR Gate for ISIDA Retriever. Установка и настройка	2	9
Принадлежность			
Код	*Обозначение*	*Версия*	
3871	*is.retriever.ocrgate*	8.*	

1. Назначение документа

Документ содержит общие указания по установке и настройке программного продукта OCR Gate for ISIDA Retriever (далее - OCRGate). OCRGate предназначен для выполнения полнотекстового распознавания изображений. OCRGate является программным мостом между прикладным решением ISIDA Retriever (и прикладными решениями, разработанными на его основе) и программными продуктами, непосредственно выполняющими полнотекстового распознавание: ABBYY FineReader Engine или Tesseract.

Важно!
OCRGate работает только с программным продуктом ABBYY FineReader Engine и Tesseract.

Документ предназначен для системного администратора.

2. Технические требования

Требования к программно-аппаратному комплексу изложены в таблице 2-1.

Таблица 2-1

№ п/п	Компонент	Характеристики/Версия
1.	Операционная система	Microsoft Windows x86/x64, Linux (только для Tesseract)
2.	SUN Java Development Kit	v.1.8.202
3.	Apache Tomcat	v.7 и выше
4.	ABBYY FineReader Engine	v.8,10 или 12
5.	Tesseract	v.3 и выше

3. Состав дистрибутива

Состав дистрибутива OCRGate представлен в таблице 3-1.

Дистрибутив OCRGate поставляется в архиве *ISIDA_Retriever-b*-Utils.zip*, где * - порядковый номер сборки.

Таблица 3-1

№ п/п	Компонент	Характеристики/Версия
Для ABBYY FineReader Engine		
1.	frengine-data	Бинарные компоненты OCRGate.
2.	webapps\frengine	Web-приложение, ответственное за приём файлов и возврат распознанного текста.
Для Tesseract		
3.	tessdata	Бинарные компоненты OCRGate.
4.	lib\tesseract.properties	Файл настроек Tesseract.
5.	webapps\tesseract	Web-приложение, ответственное за приём файлов и возврат распознанного текста.

4. Установка и настройка

4.1. Использование ABBYY FineReader Engine

Предполагается, что перед установкой OCRGate были установлены внешние (по отношению к OCRGate) программные продукты:

1. ABBYY FineReader Engine.
2. SUN Java Development Kit.
3. Apache Tomcat.

Важно!
Процедура установки и настройки внешних к OCRGate программных продуктов не описывается в данном документе и является частью сопроводительной документации этих программных продуктов.

Процедура установки и настройки OCRGate состоит из следующих шагов:

1. **Распаковка дистрибутива.**
   Дистрибутив распаковывается в корневой каталог Apache Tomcat. В итоге должна получиться следующая структура каталогов (может незначительно отличаться):

   bin\
   conf\
   frengine-data\
   bin\
   logs\
   recognizedfiles\
   temp\
   template\
   upload\

lib
logs
temp
webapps
frengine
work

2. **Настройка бинарных компонентов.**
   Все настройки OCRGate хранятся в системном реестре операционной системы. Настройка производится путём правки файла
   frengine-data\bin\frengine-x64.reg
   или
   frengine-data\bin\frengine-x86.reg
   в зависимости от используемой операционной системы. Правка производится в любом текстовом редакторе. Затем файл импортируется в системный реестр.

Параметры OCRGate, вынесенные в системный реестр операционной системы, описаны в таблице 4.1-1.

Таблица 4.1-1

№ п/п	Параметр	Описание	Пример значения
1.	ENGINE_HOME	Каталог, в который был остановлен ABBYY FineReader Engine.	c:\Program Files (x86)\ABBYY SDK\10\FineReader Engine
2.	SPOOL_DIR	Каталог файлов-результатов работы OCRGate.	c:\\Programs\\Apache Tomcat Retriever\\frengine-data\\temp
3.	ENGINE_LOG_FILE	Путь к файлу диагностики OCRGate.	c:\\Programs\\Apache Tomcat Retriever\\frengine-data\\logs\\FREngine.log
4.	ENGINE_SERIAL_NUMBER	Лицензионный ключ разработчика ABBYY FineReader Engine. Внимание! Это не лицензионный ключ конечного потребителя. При использовании ABBYY FineReader Engine версии 10 значение параметра менять не нужно.	SWRD-1010-0002-0138-5177-1128
5.	ENGINE_RECOGNIZE_LANG	Список распознаваемых языков. Через запятую перечисляются идентификаторы используемых словарей.	Russian,English
6.	ENGINE_RECOGNIZE_FORMAT	Формат результата распознавания по умолчанию. Значение «4» - plain text. Менять не рекомендуется.	4
7.	ENGINE_TMP_DIR	Каталог временных файлов OCRGate.	c:\\Programs\\Apache Tomcat Retriever\\frengine-data\\temp
8.	ENGINE_FULL_LOGGING	Включение/отключение диагностики OCRGate.	TRUE

Продолжение таблицы 4.1-1

№ п/п	Параметр	Описание	Пример значения
9.	IS_FORM_READER	Включение/отключение распознавания форм. Имеет смысл только при использовании технологии © ABBYY FormReader.	FALSE
10.	PATH_TEMPLATE	Каталог шаблонов форм. Имеет смысл только при использовании технологии © ABBYY FormReader.	c:\\Programs\\Apache Tomcat Retriever\\frengine-data\\template\\

После завершения правки файла настроек его необходимо импортировать в системный реестр операционной системы.

3. **Настройка web-приложения.**
Настройка web-приложения OCRGate сводится к указанию пути на каталог с бинарными компонентами. Это осуществляется любым текстовым редактором в файле webapps\frengine\WEB-INF\web.xml. В строке 23, значением параметра *frengine.dss.dir*, необходимо указать полный путь к каталогу с бинарными компонентами OCRGate.

К примеру:

```
<env-entry>
        <description>Directory used to temporarily store files for frengine proccess</description>
        <env-entry-name>frengine.dss.dir</env-entry-name>
        <env-entry-type>java.lang.String</env-entry-type>
        <env-entry-value>c:\\Programs\\Apache Tomcat Retriever\\frengine-data</env-entry-value>
</env-entry>
```

4.2. Использование Tesseract

Предполагается, что перед установкой OCRGate были установлены внешние (по отношению к OCRGate) программные продукты:

4. Tesseract (в окружении Linux).
5. SUN Java Development Kit.
6. Apache Tomcat.

Важно!
Процедура установки и настройки внешних к OCRGate программных продуктов не описывается в данном документе и является частью сопроводительной документации этих программных продуктов.

Процедура установки и настройки OCRGate состоит из следующих шагов:

1. **Распаковка дистрибутива.**
Дистрибутив распаковывается в корневой каталог Apache Tomcat. В итоге должна получиться следующая структура каталогов (может незначительно отличаться):

```
bin\
conf\
tesseract\
        tessdata\
                *\
lib\
        tesseract.properties
logs\
temp\
webapps\
        tesseract\
work\
```

2. Настройка.

Все настройки OCRGate хранятся в файле lib\tesseract.properties. Правка файла производится в любом текстовом редакторе.

Параметры OCRGate для Tesseract описаны в таблице 4.2-1.

Таблица 4.2-1

№ п/п	Параметр	Описание	Пример значения
1.	engine.dss.dir	Путь к каталогу tessdata. Сам каталог tessdata в пути не указывается.	/opt/tesseract
2.	engine.language.list	Список используемых языков. Разделитель - запятая.	RUS,ENG
3.	engine.logging	Включение/отключение отладочной диагностики.	false

5. Проверка работоспособности

Работоспособность OCRGate проверяется с помощью специализированной web-страницы приложения frengine.

Для проверки следует, с помощью любого web-браузера, обратиться к приложению frengine, указать любой графический файл, содержащий текст, и отправить его на распознавание. В случае успешно выполненного распознавания приложением будет возвращён распознанный текст.