

[Docs](#) » Welcome to FSCrawler's documentation!

Welcome to FSCrawler's documentation!

⚠ Warning

This documentation is for the version of FSCrawler currently under development. Were you looking for the [documentation of the latest stable version](#)?

Welcome to the FS Crawler for [Elasticsearch](#).

This crawler helps to index binary documents such as PDF, Open Office, MS Office.

Main features:

- Local file system (or a mounted drive) crawling and index new files, update existing ones and removes old ones.
- Remote file system over SSH crawling.
- REST interface to let you “upload” your binary documents to elasticsearch.

ⓘ Note

FS Crawler 2.7-SNAPSHOT is using [Tika 1.22](#) and:

- [Elasticsearch Rest Client 7.3.0](#) for Elasticsearch V7.
- [Elasticsearch Rest Client 6.8.1](#) for Elasticsearch V6.
- [Elasticsearch Rest Client 5.6.15](#) for Elasticsearch V5.

Installation Guide

- [Download FSCrawler](#)
- [Running as a Service on Windows](#)
- [Upgrade FSCrawler](#)
 - [Upgrade to 2.2](#)
 - [Upgrade to 2.3](#)
 - [Upgrade to 2.4](#)
 - [Upgrade to 2.5](#)
 - [Upgrade to 2.6](#)
 - [Upgrade to 2.7](#)

- [Getting Started](#)
 - [Start FSCrawler](#)
 - [Searching for docs](#)
 - [Ignoring folders](#)
- [Crawler options](#)
- [OCR integration](#)
 - [OCR settings](#)
 - [Disable/Enable OCR](#)
 - [OCR Language](#)
 - [OCR Path](#)
 - [OCR Data Path](#)
 - [OCR Output Type](#)
 - [OCR PDF Strategy](#)
- [Starting with a REST gateway](#)
- [Supported formats](#)
- [Tips and tricks](#)
 - [Moving files to a “watched” directory](#)
 - [Indexing from HDFS drive](#)
 - [Using docker](#)

Administration Guide

- [Status files](#)
- [CLI options](#)
 - [Upgrade](#)
 - [Loop](#)
 - [Restart](#)
 - [Rest](#)
- [JVM Settings](#)
- [Configuring an external logger configuration file](#)
- [Job file specification](#)
- [The most simple crawler](#)
- [Local FS settings](#)
 - [Root directory](#)
 - [Update rate](#)
 - [Includes and excludes](#)
 - [Filter content](#)
 - [Indexing JSon docs](#)
 - [Indexing XML docs](#)
 - [Add as Inner Object](#)

- [Index folders](#)
- [Dealing with multiple types and multiple dirs](#)
- [Dealing with multiple types within the same dir](#)
- [Using filename as elasticsearch `_id`](#)
- [Adding file attributes](#)
- [Disabling raw metadata](#)
- [Disabling file size field](#)
- [Ignore deleted files](#)
- [Ignore content](#)
- [Continue on Error](#)
- [Language detection](#)
- [Storing binary source document](#)
- [Extracted characters](#)
- [Ignore Above](#)
- [File checksum](#)
- [Follow Symlinks](#)
- [SSH settings](#)
 - [Username / Password](#)
 - [Using Username / PEM file](#)
- [Elasticsearch settings](#)
 - [Index settings](#)
 - [Index settings for documents](#)
 - [Index settings for folders](#)
 - [Mappings](#)
 - [Bulk settings](#)
 - [Using Ingest Node Pipeline](#)
 - [Node settings](#)
 - [Using Credentials \(X-Pack\)](#)
 - [SSL Configuration](#)
 - [Generated fields](#)
 - [Search examples](#)
- [REST service](#)
 - [FSCrawler status](#)
 - [Uploading a binary document](#)
 - [Simulate Upload](#)
 - [Document ID](#)
 - [Additional tags](#)
 - [Specifying an elasticsearch index](#)
 - [Enabling CORS](#)
 - [REST settings](#)

Developer Guide

- [Building the project](#)
 - [Clone the project](#)
 - [Build the artifact](#)
 - [Integration tests](#)
 - [Run tests from your IDE](#)
 - [Run tests with an external cluster](#)
 - [Using security feature](#)
 - [Tests options](#)
 - [Check for vulnerabilities \(CVE\)](#)
- [Writing documentation](#)
- [Release the project](#)

License

! Important

This software is **licensed** under the Apache 2 **license**, quoted below.

Copyright 2011-2019 David Pilato

Licensed under the Apache **License**, Version 2.0 (the “**License**”); you may not use this file except in compliance with the **License**. You may obtain a copy of the **License** at

<http://www.apache.org/licenses/LICENSE-2.0>

Unless required by applicable law or agreed to in writing, software distributed under the **License** is distributed on an “AS IS” BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the **License** for the specific language governing permissions and limitations under the **License**.

Incompatible 3rd party library licenses

Some libraries are not Apache2 compatible. Therefore they are not packaged with FSCrawler so you need to download and add manually them to the `lib` directory:

- for JBIG2 images, you need to add [levigo-jbig2-imageio:2.0](#) library
- for TIFF images, you need to add [jai-imageio-core:1.4.0](#) library
- for JPEG 2000 (JPX) images, you need to add [jai-imageio-jpeg2000:1.3.0](#) library

See [pdfbox documentation](#) for more details.

Special thanks

Thanks to [JetBrains](#) for the IntelliJ IDEA License!

