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

User Guide

Administration Guide

Developer Guide

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!