



# Indexing your office documents with Elastic and FSCrawler

David Pilato  
*Developer / Evangelist, Community*  
[@dadoonet](#)



## Apache Tika - a content analysis toolkit

The Apache Tika™ toolkit detects and extracts metadata and text from over a thousand different file types (such as PPT, XLS, and PDF). All of these file types can be parsed through a single interface, making Tika useful for search engine indexing, content analysis, translation, and much more. You can find the latest release on the [download page](#). Please see the [Getting Started](#) page for more information on how to start using Tika.

The [Parser](#) and [Detector](#) pages describe the main interfaces of Tika and how they work.

If you're interested in contributing to Tika, please see the [Contributing](#) page or send an email to the [Tika development list](#).

Tika is a project of the [Apache Software Foundation](#), and was formerly a subproject of [Apache Lucene](#).

## Latest News

### 18 August 2021: Apache Tika Release

Apache Tika 2.1.0 has been released! This release includes a significant refactoring in the tika-parsers-extended modules and numerous bug fixes (especially in the pipes modules) and dependency upgrades. Please see the [CHANGES.txt](#) file for the full list of changes in the release and have a look at the download page for more information on how to obtain Apache Tika 2.1.0.

### 19 July 2021: Apache Tika Release

Apache Tika 2.0.0 has been released! This release includes a significant refactoring from the 1.x branch including modularization of the parsers modules, the new pipes modules and numerous bug fixes and dependency upgrades. Please see the [CHANGES.txt](#) file for the full list of changes in the release and have a look at the download page for more information on how to obtain Apache Tika 2.0.0.

### 6 July 2021: Apache Tika Release

Apache Tika 1.27 has been released! This release includes a new JSON handler for the /tika endpoint in tika-server, a new MP4 parser based on Drew Noakes' metadata-extractor and numerous bug fixes and dependency upgrades. Please see the [CHANGES.txt](#) file for the full list of changes in the release and have a look at the download page for more information on how to obtain Apache Tika 1.27.

### Apache Tika

[Introduction](#)  
[Download](#)  
[Contribute](#)  
[Mailing Lists](#)  
[Tika Wiki](#)  
[Issue Tracker](#)  
[Security](#)

### Documentation

- [Apache Tika 2.1.0](#)
  - [Getting Started](#)
  - [Supported Formats](#)
  - [Parser API](#)
  - [Parser 5min Quick Start Guide](#)
  - [Content and Language Detection](#)
  - [Configuring Tika](#)
  - [Usage Examples](#)
  - [API Documentation](#)
  - [REST API Documentation \(Miredot\)](#)
- [Apache Tika 2.0.0](#)
- [Apache Tika 1.27](#)
- [Apache Tika 1.26](#)
- [Apache Tika 1.25](#)
- [Apache Tika 1.24.1](#)
- [Apache Tika 1.24](#)
- [Apache Tika 1.23](#)
- [Apache Tika 1.22](#)
- [Apache Tika 1.21](#)
- [Apache Tika 1.20](#)
- [Apache Tika 1.19.1](#)
- [Apache Tika 1.19](#)
- [Apache Tika 1.18](#)
- [Apache Tika 1.17](#)
- [Apache Tika 1.16](#)
- [Apache Tika 1.15](#)
- [Apache Tika 1.14](#)
- [Apache Tika 1.13](#)
- [Apache Tika 1.12](#)
- [Apache Tika 1.11](#)
- [Apache Tika 1.10](#)
- [Apache Tika 1.9](#)
- [Apache Tika 1.8](#)
- [Apache Tika 1.7](#)
- [Apache Tika 1.6](#)
- [Apache Tika 1.5](#)
- [Apache Tika 1.4](#)
- [Apache Tika 1.3](#)
- [Apache Tika 1.2](#)
- [Apache Tika 1.1](#)
- [Apache Tika 1.0](#)

**Please note** that Apache Tika is able to detect a much wider range of formats than those listed below, this page only documents those formats from which Tika is able to extract metadata and/or textual content.

- [Supported Document Formats](#)
  - [HyperText Markup Language](#)
  - [XML and derived formats](#)
  - [Microsoft Office document formats](#)
  - [OpenDocument Format](#)
  - [iWorks document formats](#)
  - [WordPerfect document formats](#)
  - [Portable Document Format](#)
  - [Electronic Publication Format](#)
  - [Rich Text Format](#)
  - [Compression and packaging formats](#)
  - [Text formats](#)
  - [Feed and Syndication formats](#)
  - [Help formats](#)
  - [Audio formats](#)
  - [Image formats](#)
  - [Video formats](#)
  - [Java class files and archives](#)
  - [Source code](#)
  - [Mail formats](#)
  - [CAD formats](#)
  - [Font formats](#)
  - [Scientific formats](#)
  - [Executable programs and libraries](#)
  - [Crypto formats](#)
  - [Database formats](#)
  - [Natural Language Processing](#)
  - [Image and Video object recognition](#)

# Parsing a stream

## and getting content and metadata

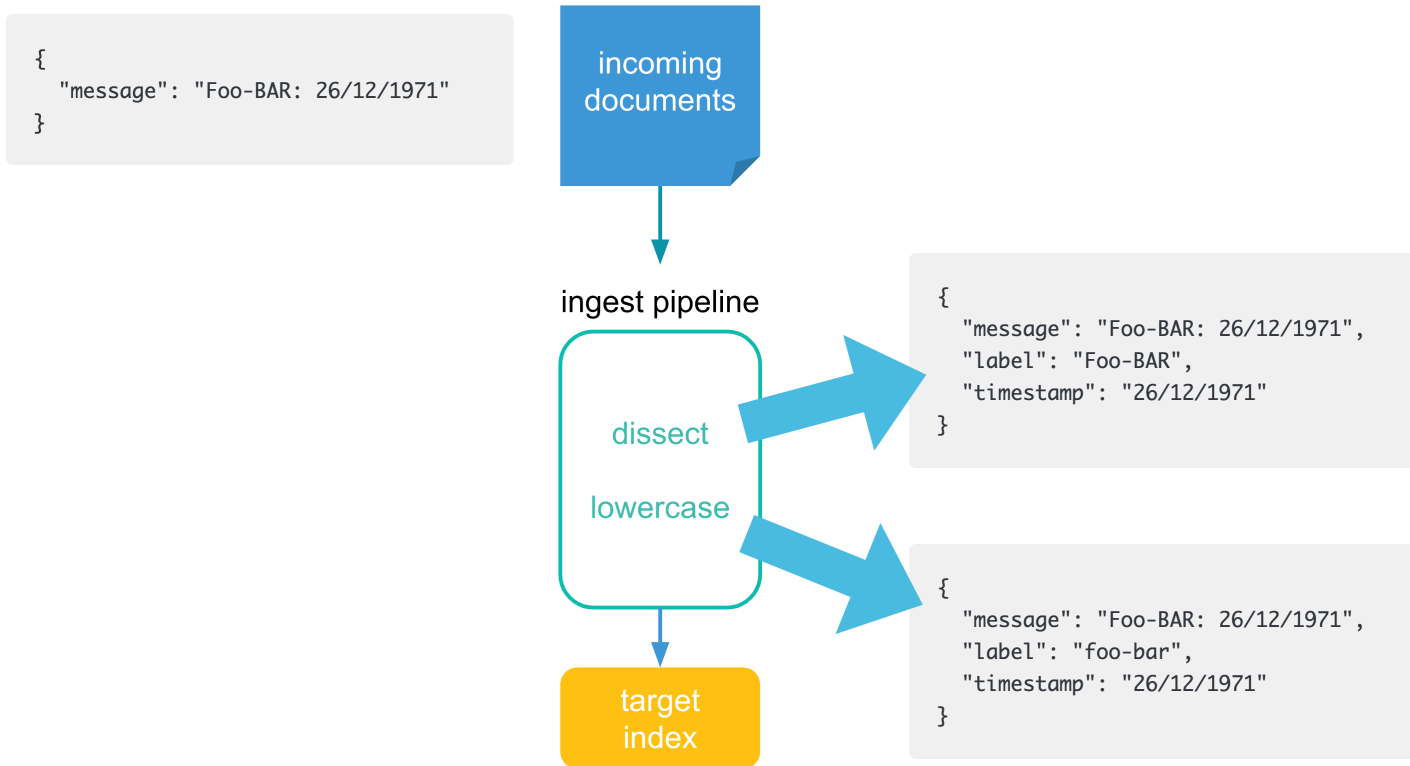
```
static void extractTextAndMetadata(InputStream stream) throws Exception {
    BodyContentHandler handler = new BodyContentHandler();
    Metadata metadata = new Metadata();
    try (stream) {
        new DefaultParser().parse(stream, handler, metadata, new ParseContext());
        String extractedText = handler.toString();
        String title = metadata.get(TikaCoreProperties.TITLE);
        String keywords = metadata.get(TikaCoreProperties.KEYWORDS);
        String author = metadata.get(TikaCoreProperties.CREATOR);
    }
}
```



# **ingest-attachment plugin** extracting from BASE64 or CBOR




# An ingest pipeline



# ingest-attachment processor plugin


using Tika behind the scene

 [Products](#) [Customers](#) [Learn](#) [Company](#) [Pricing](#)

**Docs**  
[Elasticsearch Plugins and Integrations \[7.12\]](#) » [Ingest Plugins](#) » **Ingest Attachment Processor Plugin**

[« Ingest Plugins](#) [Using the Attachment Processor in a Pipeline »](#)

## Ingest Attachment Processor Plugin




The ingest attachment plugin lets Elasticsearch extract file attachments in common formats (such as PPT, XLS, and PDF) by using the Apache text extraction library [Tika](#).

You can use the ingest attachment plugin as a replacement for the mapper attachment plugin.

The source field must be a base64 encoded binary. If you do not want to incur the overhead of converting back and forth between base64, you can use the CBOR format instead of JSON and specify the field as a bytes array instead of a string representation. The processor will skip the base64 decoding then.

### Installation



This plugin can be installed using the plugin manager:

```
sudo bin/elasticsearch-plugin install ingest-attachment
```



# Demo



<https://cloud.elastic.co>

SEDONA








# FSCrawler

You know, for files...





Pull requests
Issues
Marketplace
Explore

dadoonet / fscrawler
Unwatch 76
Unstar 812
Fork 203

<> Code
Issues 102
Pull requests 8
Actions
Projects 2
Security
Insights
Settings

master
15 branches
19 tags
Go to file
Add file
Code

**mergify** Merge pull request #997 from dadoonet/dependabot/mave...
 4669ef2 20 days ago
1,217 commits

.github	Update the issue templates	4 months ago
.mvn	Move to .mvn folder all needed settings to build/test FSCrawler	4 years ago
beans	Add support for YAML configuration	2 years ago
cli	Remove support for Elasticsearch v5	9 months ago
contrib/docker-compose-example	Update Dockerfile-fscrawler	29 days ago
core	Fix SSH crawling from Windows machine	2 months ago
crawler	Add documentation about Windows drives SSH indexing	6 months ago
distribution	Remove support for Elasticsearch v5	9 months ago
docs	Updated documentation for instructions on how to use the contri...	2 months ago
elasticsearch-client	Add `path_prefix` option	6 months ago
framework	Remove support for Elasticsearch v5	9 months ago
integration-tests	Fix flaky tests	2 months ago
rest	Add more information to the _simulate API	9 months ago
settings	Document `auto` option for `pdf_strategy`	3 months ago
src/main/resources/org/apache/...	Have tests for ES5 and ES6 in the same repo (no more profiles)	2 years ago
test-documents	Document `auto` option for `pdf_strategy`	3 months ago

### About

Elasticsearch File System Crawler (FS Crawler)

[fscrawler.readthedocs.io/](https://fscrawler.readthedocs.io/)

[java](#)
[elasticsearch](#)
[crawler](#)
[tika](#)

Readme
 Apache-2.0 License

---

### Releases 19

**FSCrawler 2.6** Latest  
 on 9 Jan 2019

[+ 18 releases](#)

---


### Packages

No packages published

[Publish your first package](#)

---

### Used by 7





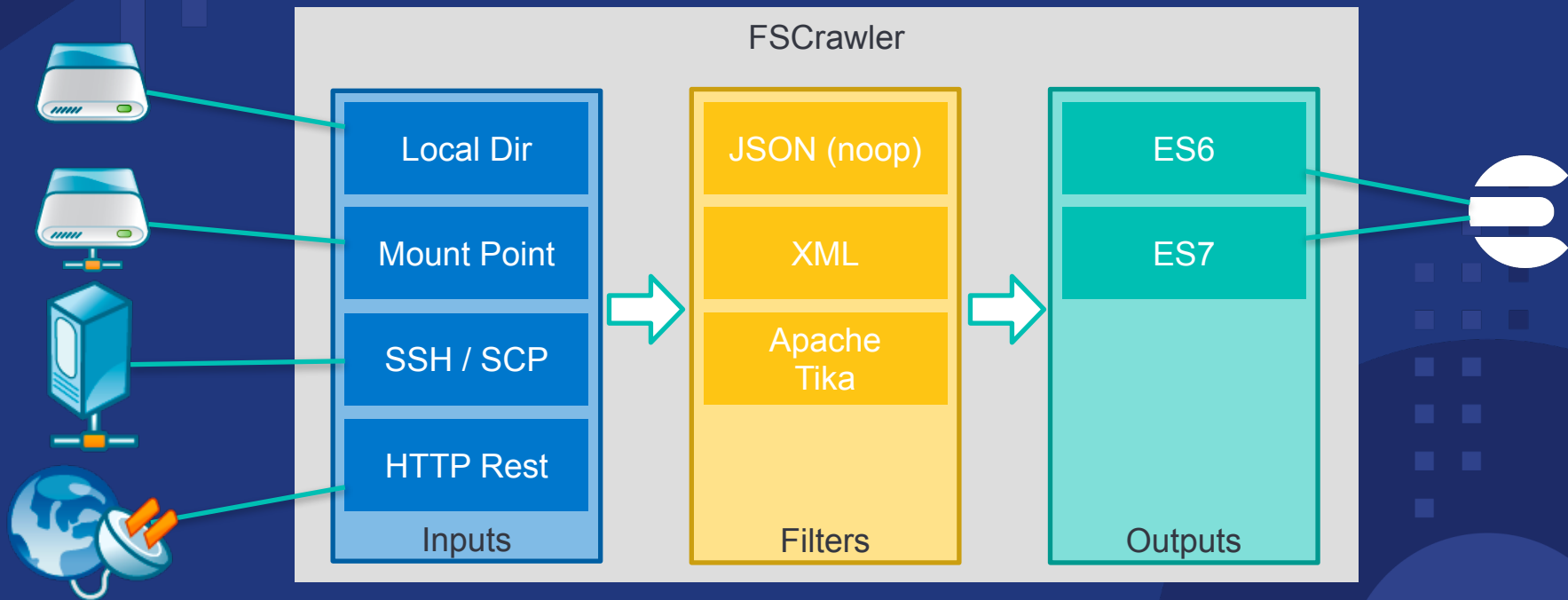
# *Disclaimer*

This project is a community project.  
**It is not officially supported by Elastic.**  
Support is only provided by FSCrawler community  
on discuss and stackoverflow.

<http://discuss.elastic.co/>  
<https://stackoverflow.com/questions/tagged/fscrawler>

# FSCrawler

## Architecture



# FSCrawler

## Key Features

- Much more formats than ingest attachment plugin
- OCR (Tesseract)
- Much more metadata than ingest attachment plugin  
(See <https://fscrawler.readthedocs.io/en/latest/admin/fs/elasticsearch.html#generated-fields>)
- Language detection

# Documentation

- <https://fscrawler.readthedocs.io/>
- <https://fscrawler.readthedocs.io/en/latest/user/tutorial.html>
- <https://fscrawler.readthedocs.io/en/latest/user/formats.html>
- <https://fscrawler.readthedocs.io/en/latest/admin/fs/index.html>



# FSCrawler



even better with a UI



# FSCrawler

Workplace Search integration

## Add Workplace Search connector #991

 Merged  dadoonet merged 82 commits into `master` from `wip/workplace_search` on 22 Dec 2020

 Conversation 3

 Commits 82

 Checks 5

 Files changed 106



**dadoonet** commented on 30 Jul 2020 · edited ▾

Owner 😊 ⋮

This PR adds a connector to Workplace Search.

### Setup

Full documentation available at: [https://fscrawler.readthedocs.io/en/wip-workplace\\_search/admin/fs/wpsearch.html](https://fscrawler.readthedocs.io/en/wip-workplace_search/admin/fs/wpsearch.html)

### Keys

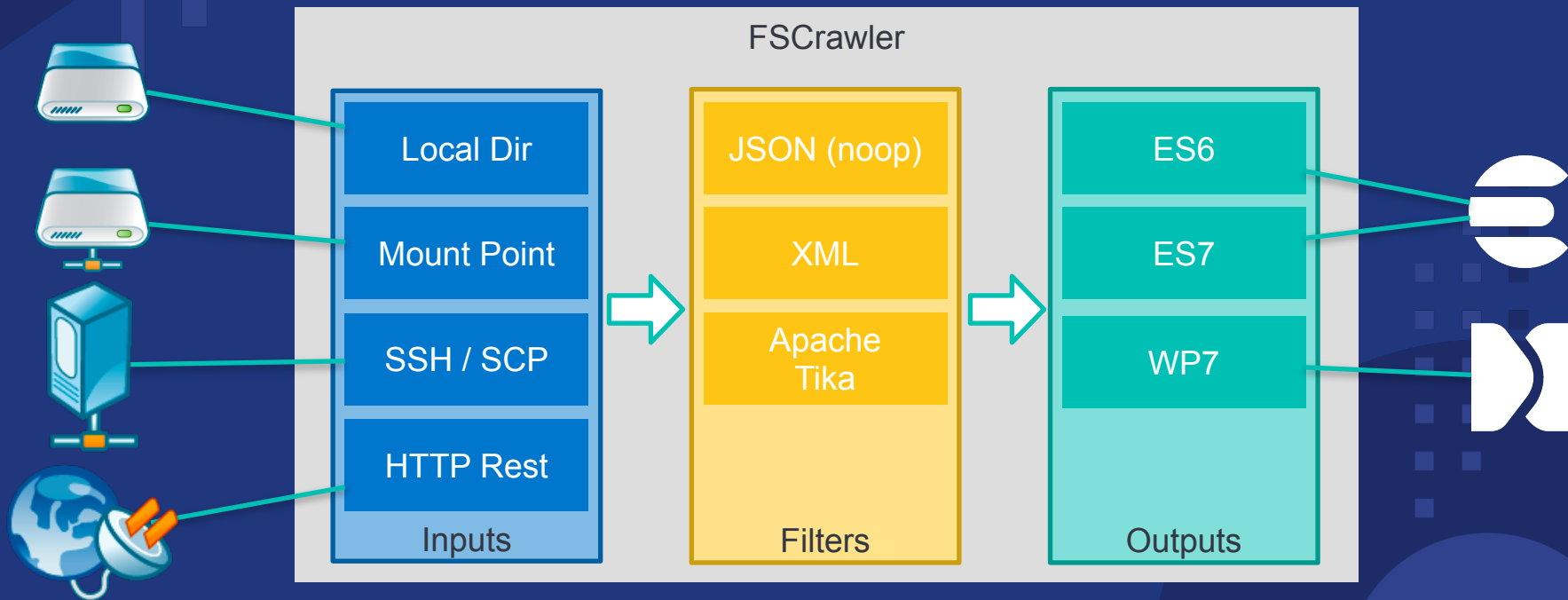
Once you have created your Custom API and have the `ACCESS_TOKEN` and `KEY`, you can add to your existing FSCrawler configuration file:

```
name: "test"
workplace_search:
```



# FSCrawler

## Architecture





# Demo



<https://cloud.elastic.co>





# Thanks!

PR are warmly welcomed!

<https://github.com/dadoonet/fscrawler>