Tika2_0MigrationGuide

Tika 2.0 Migration Guide

This page dynamically documents changes that Tika users should be aware of when they migrate to Tika 2.0. While Tika 2.0 is still very much in the early development stages, it will be helpful to gather breaking changes and other important modifications here.

This page differs from Tika2_0RoadMap in that this page documents changes that have been made to the 2.x branch, while the road map documents potential changes.

Major Changes

Tika Modules

In Tika 2.x the tika-parsers project has been split into 15 separate modules. With Tika's ever growing list of parsers the modules give developers the ability to pick and choose sub-groupings of parsers without bringing every parser dependency into a project. For example projects using Tika 1.12 parsers would include the following entry in an Apache Maven pom.xml dependency element:

```
<dependency>
    <groupId>org.apache.tika</groupId>
    <artifactId>tika-parsers</artifactId>
    <version>1.12</version>
</dependency>
```

If this project were only being used to parse PDF files this could be refactored to the entry below on Tika 2.x:

```
<dependency>
    <groupId>org.apache.tika</groupId>
    <artifactId>tika-parser-pdf-module</artifactId>
    <version>2.0</version>
</dependency>
```

The 2.x branch also introduces the ParserProxy, DetectorProxy, and EncodingDetectorProxy classes that allow developers to compose Parsers, Detectors and EncodingDetectors using classes that may or may not exist on the classpath.

For example the OutlookExtractor exists in the tika-parser-office-module. An Outlook message may contain HTML content but the user may not want to include the tika-parser-web-module that contains the HtmlParser. By wrapping the HtmlParser in a ParserProxy:

```
this.htmlParserProxy = new ParserProxy("org.apache.tika.parser.html.HtmlParser", getClass().getClassLoader());
```

The OutlookExtractor will only parse the HTML content if that module is included. If not the parser fails silently by default. In cases where the developer wants to be warned that a proxy has failed the developer may set the following System Property:

```
org.apache.tika.service.proxy.error.warn=true
```

Which will print a warning message when the class being proxied is not found.

Tika Bundles

In addition to replacing tika-parsers the tika-bundle will also be replaced in Tika 2.x. Instead of a single tika-bundle there will be a bundle for each parser module created above. For example tika-parser-pdf-module will have a corresponding tika-parser-pdf-bundle. As in tika-bundle the dependencies in the bundle projects will be embedded in the jar file to allow OSGi unfriendly projects to be easily included. The bundles will also start up an OSGi service for each listed service in the META-INF/services/ file entries. This will allow OSGi developers an easier way to get access to individual parsers through the OSGi service registry. Finally a TikaService class will also be added to the OSGi Service registry that will serve as a means to put a document through all the parsers available in the OSGi service registry.

Minor Changes