

TikaExtractingEmbeddedCode

Extracting Embedded VBA and JS

By default, Tika ignores embedded VBA and js. The user must configure this programmatically or via tika-config.xml:

```
<?xml version="1.0" encoding="UTF-8"?>
<properties>
  <parsers>
    <parser class="org.apache.tika.parser.DefaultParser">
      <parser-exclude class="org.apache.tika.parser.html.HtmlParser"/>
      <parser-exclude class="org.apache.tika.parser.pdf.PDFParser"/>
      <parser-exclude class="org.apache.tika.parser.microsoft.ooxml.OOXMLParser"/>
      <parser-exclude class="org.apache.tika.parser.microsoft.OfficeParser"/>
    </parser>

    <parser class="org.apache.tika.parser.html.HtmlParser">
      <params>
        <param name="extractScripts" type="bool">true</param>
      </params>
    </parser>
    <parser class="org.apache.tika.parser.pdf.PDFParser">
      <params>
        <param name="extractActions" type="bool">true</param>
      </params>
    </parser>
    <parser class="org.apache.tika.parser.microsoft.ooxml.OOXMLParser">
      <params>
        <param name="extractMacros" type="bool">true</param>
      </params>
    </parser>
    <parser class="org.apache.tika.parser.microsoft.OfficeParser">
      <params>
        <param name="extractMacros" type="bool">true</param>
      </params>
    </parser>
  </parsers>
</properties>
```

We encourage using the `RecursiveParserWrapper` for easier understanding of the extracted data and the boundaries between the parent file and the embedded files – the `-J` option in `tika-app` or the `/rmeta` endpoint in `tika-server`.