TikaPlugin

Tika Plugin

The Tika plugin in http://issues.apache.org/jira/browse/NUTCH-766 is a first attempt at delegating the parsing to Tika instead of having to maintain the parser plugins in Nutch. This page will list the differences in coverage or functionality between the Tika plugin and the existing Nutch parsers. Tika also has more formats not covered by Nutch which are not described here and has a more generic capability of representing structured content which can be useful for HtmlParseFilters (which are currently limited to HTML content).

html: comparable

js: ?

mp3: Nutch identifies several fields (Title, Album, Artist) whereas Tika knows only about Titles, the rest is stored as paragraphs.

Tika-app can also identify in an mp3 id3v1 and id3v2 tags like: album, artist, audioSampleRate, composer, genre, logcomment, releaseDate, trackNumber using the XMPDM interface

msexcel: comparable (+ Tika able to represent content in structured way as XHTML tables which can be useful for HTML parser plugins)

mspowerpoint: comparable

msword: Tika does not support word 95 other versions are comparable

openoffice: comparable

pdf: comparable

rss: Tika identifies only the Mimetype but does nothing about the content

rtf: deactivated in Nutch for licensing reasons | works in Tika

swf : not yet covered in Tika (see https://issues.apache.org/jira/browse/TIKA-337)

text: comparable

zip: ?