[Jeroen Reijn](http://blog.jeroenreijn.com/)

About Java and software development

**05 APRIL 2010**

Metadata extraction with Apache Tika

[http://1.bp.blogspot.com/_hd6Y7yyFK7E/S7pGYKIkWSI/AAAAAAAAAX4/LLz_aNEujsM/s320/tika.png](http://1.bp.blogspot.com/_hd6Y7yyFK7E/S7pGYKIkWSI/AAAAAAAAAX4/LLz_aNEujsM/s1600/tika.png)

At [Hippo](http://www.onehippo.com/) I work with/for customers that have quite a lot of content. The projects I work on have content in the range of 5.000 to 500.000 document gathered in one content repository. This can be just textual content, but most of the time this is a variety of different content types. You might think of images, PDFs and Microsoft office document formats. By default [Apache JackRabbit](http://jackrabbit.apache.org/), the layer underneath [Hippo Repository](http://docs.onehippo.org/), indexes this kind of content by using extractors, so that the information can be found within the [Hippo CMS 7](http://docs.onehippo.org/) search or from any application connected to the Hippo Repository which is performing a search on the content repository. Being able to search on content found within a file is interesting, but there is so much more that you can do with this kind of information.

**CONTENT METADATA**

Having all this content inside the repository is nice, but a certain piece of content uploaded to the repository can contain much more information than the file itself and sometimes this metadata is ignored in content management systems. As an example you might want to be able to see the number of pages of a PDF document inside your CMS or view the [EXIF](http://en.wikipedia.org/wiki/Exchangeable_image_file_format) information of an image stored inside your content repository. There are a number of parser libraries out there that can extract information from a specific file format, but you can get quite lost. Within the ASF there is also a very nice toolkit called [Apache Tika](http://lucene.apache.org/tika/), which provides parsers for a lot of different file formats.

**WHAT IS APACHE TIKA?**

Apache Tika is a subproject of the [Apache Lucene](http://lucene.apache.org/) project and is a toolkit for extracting content and metadata from different kind of file formats. The content extraction logic is not located inside Tika itself, but Tika defines a standard API and makes use of existing libraries like [POI](http://poi.apache.org/) and [PDFBox](http://pdfbox.apache.org/) for it's content extraction. While writing this post the current release of Tika is version 0.6 and the following file formats are already supported:

* HyperText Markup Language
* XML and derived formats
* Microsoft Office document formats
* OpenDocument Format
* Portable Document Format
* Electronic Publication Format
* Rich Text Format
* Compression and packaging formats
* Text formats
* Audio formats
* Image formats
* Video formats
* Java class files and archives
* The mbox format

As you can see this is already quite a lot. The team behind Tika is working hard on improving the current parser possibilities and adding more formats for the upcoming releases. Tika is actually already being used by a number of other Apache projects like JackRabbit and [Solr](http://lucene.apache.org/solr/). Now let's see how we can use Tika ourselves.

**GETTING STARTED**

I always work with [Maven](http://maven.apache.org/) as my build system, so let's start of with a piece of pom.xml. First add the Tika parser dependency to our pom.xml.

[?](http://blog.jeroenreijn.com/2010/04/metadata-extraction-with-apache-tika.html)

|  |  |
| --- | --- |
| 1  2  3  4  5 | <dependency>    <groupId>org.apache.tika</groupId>    <artifactId>tika-parsers</artifactId>    <version>0.6</version>  </dependency> |

By depending on *tika-parsers* Maven will automatically gather the required parser libraries, which are needed to parse certain file formats. Since my Java code example will be based on a unit test, we will also need to add JUnit as a dependency to our pom.xml.

[?](http://blog.jeroenreijn.com/2010/04/metadata-extraction-with-apache-tika.html)

|  |  |
| --- | --- |
| 1  2  3  4  5 | <dependency>    <groupId>junit</groupId>    <artifactId>junit</artifactId>    <version>4.7</version>  </dependency> |

In this post I want to see what kind of [EXIF](http://en.wikipedia.org/wiki/Exchangeable_image_file_format) information can be retrieved from an image by using Tika. One of my hobbies is photography and therefor I have tons of images, which contain a lot of metadata about for instance the ISO speed or dimensions of an image. Now let's write some actual code to see how Tika works.

**THE ACTUAL CODE**

As I mentioned before, my example code is written as a [JUnit](http://www.junit.org/) test. If you are not familiar with writing tests or JUnit itself please have a look at the JUnit [website](http://www.junit.org/).  
To be able to run this test, I've added one of my images on the test classpath, so my test class will be able to find the image resource. The following piece of code shows my entire test class.

[?](http://blog.jeroenreijn.com/2010/04/metadata-extraction-with-apache-tika.html)

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38 | public class ImageMetaDataTest {        private static final String fileName = "IMG\_2659.JPG";        private Tika tika;      private InputStream stream;        @Before      public void setUp() {          tika = new Tika();          stream = this.getClass().getResourceAsStream(fileName);      }        @Test      public void testImageMetadataCameraModel() throws IOException, SAXException,              TikaException {            Metadata metadata = new Metadata();          ContentHandler handler = new DefaultHandler();          Parser parser = new JpegParser();          ParseContext context = new ParseContext();            String mimeType = tika.detect(stream);          metadata.set(Metadata.CONTENT\_TYPE, mimeType);            parser.parse(stream,handler,metadata,context);          assertTrue("The expected Model is not correct",                  metadata.get("Model").equals("Canon EOS 350D DIGITAL"));      }        @After      public void close() throws IOException {          if(stream!=null) {              stream.close();          }      }    } |

As you can see the code is quite small. The most important part of the above code example is using the JpegParser to parse the .JPG file and the creation of the Metadata object with the appropriate information.  
I think this simple test case shows you how easy to use the Tika API is. Of course in the above test case I only test for the current Camera Model, but the Metadata object holds much more information then just that.  
  
Viewing all the fields found in the metadata of the image can be achieved quite easily by using for instance the following method.

[?](http://blog.jeroenreijn.com/2010/04/metadata-extraction-with-apache-tika.html)

|  |  |
| --- | --- |
| 1  2  3  4  5  6 | private void listAvailableMetaDataFields(final Metadata metadata) {    for(int i = 0; i <metadata.names().length; i++) {      String name = metadata.names()[i];      System.out.println(name + " : " + metadata.get(name));    }  } |

The output of this method can be like:  
  
  
Easy Shooting Mode : Manual  
Image Type : Canon EOS 350D DIGITAL  
Model : Canon EOS 350D DIGITAL  
Metering Mode : Evaluative  
Quality : Fine  
Shutter/Auto Exposure-lock Buttons : AF/AE lock  
ISO Speed Ratings : 400  
  
  
It's as easy as that.

**LOOKING AHEAD**

I'm currently looking for possibilities of integrating Apache Tika into Hippo CMS 7, to enhance the system with much more metadata then there currently is available in the system. I think this can become quite a powerful addition in combination with the facetted navigation feature introduced in [Hippo CMS 7.3](http://www.onehippo.com/en/news/2010/03/CMS+7.3.html). I've already started working on some code, which I hope to provide as a patch in the near future.

Geplaatst door Jeroen Reijn op [10:59 PM](http://blog.jeroenreijn.com/2010/04/metadata-extraction-with-apache-tika.html)

[Email This](http://www.blogger.com/share-post.g?blogID=2962867622850517744&postID=2556839578610686454&target=email)[BlogThis!](http://www.blogger.com/share-post.g?blogID=2962867622850517744&postID=2556839578610686454&target=blog)[Share to Twitter](http://www.blogger.com/share-post.g?blogID=2962867622850517744&postID=2556839578610686454&target=twitter)[Share to Facebook](http://www.blogger.com/share-post.g?blogID=2962867622850517744&postID=2556839578610686454&target=facebook)

 Labels: [hippo](http://blog.jeroenreijn.com/search/label/hippo), [java](http://blog.jeroenreijn.com/search/label/java)

6 comments:

1. http://img2.blogblog.com/img/b36-rounded.png

[**Madhu**](http://www.blogger.com/profile/01083038920480436512)[June 27, 2010 at 12:41 PM](http://blog.jeroenreijn.com/2010/04/metadata-extraction-with-apache-tika.html?showComment=1277635295238#c7393303224475678478)

Hi Reijn,   
  
I've a requirement wherein I'll be getting the images and the text content as part of PDF/Word document from a third party. I need to parse those files and display the image and the content on site.   
  
Currently we are doing it manually but want to automate the process. So will Tika be helpful in this case   
  
Thanks   
Madhu

[Reply](javascript:;)

1. http://img2.blogblog.com/img/b36-rounded.png

[**Laxman Rana**](http://www.blogger.com/profile/17671523461706798298)[August 22, 2012 at 6:28 AM](http://blog.jeroenreijn.com/2010/04/metadata-extraction-with-apache-tika.html?showComment=1345609707070#c6004159724667440359)

Hi,  
I want to convert html contents to word with all styles.so will tika will be helpful??  
  
Thanks

[Reply](javascript:;)

1. 

[**Jeroen Reijn**](http://www.blogger.com/profile/00364880674735561006)[August 22, 2012 at 9:01 AM](http://blog.jeroenreijn.com/2010/04/metadata-extraction-with-apache-tika.html?showComment=1345618868261#c6515410325538493406)

No Tika is only helpful when you want to detects and extract metadata. You might want to take a look at Apache POI (http://poi.apache.org/).

[Reply](javascript:;)

1. http://img2.blogblog.com/img/b36-rounded.png

[**Anil Mudalgi**](http://www.blogger.com/profile/01245525688841055797)[September 10, 2012 at 2:37 PM](http://blog.jeroenreijn.com/2010/04/metadata-extraction-with-apache-tika.html?showComment=1347280645031#c2180785609593376839)

Hi,  
  
I wanted to convert doc/docx file to html. can Tika help me. If yes can you me some code sample. I'm working on c#

[Reply](javascript:;)

[Replies](javascript:;)

* 1. 

[**Jeroen Reijn**](http://www.blogger.com/profile/00364880674735561006)[September 14, 2012 at 11:03 AM](http://blog.jeroenreijn.com/2010/04/metadata-extraction-with-apache-tika.html?showComment=1347613384798#c7800661087769709609)

Hi I'm not sure if there is anything for C#. Tika should be able to extract a document and convert it to HTML yes.

[**Reply**](javascript:;)

1. http://img1.blogblog.com/img/anon36.png

**ivanov-void**[December 11, 2012 at 3:29 PM](http://blog.jeroenreijn.com/2010/04/metadata-extraction-with-apache-tika.html?showComment=1355236170881" \l "c4921875966385835345)

very helpfull! Thanks!

[Reply](javascript:;)

[Newer Post](http://blog.jeroenreijn.com/2010/05/photography-in-artis.html" \o "Newer Post)[Older Post](http://blog.jeroenreijn.com/2010/02/jboss-modeshape-federating-jcr.html)[Home](http://blog.jeroenreijn.com/)

Subscribe to: [Post Comments (Atom)](http://blog.jeroenreijn.com/feeds/2556839578610686454/comments/default)

**PAGES**

* [Home](http://blog.jeroenreijn.com/)
* [About me](http://about.me/jeroenreijn)
* [GitHub](http://github.com/jreijn/)
* [Twitter](http://twitter.com/jreijn)

**ABOUT ME**

[](https://plus.google.com/101111641807482638731)

[**Jeroen Reijn**](https://plus.google.com/101111641807482638731)

Working as a Solution Architect at [Hippo](http://www.onehippo.com/), committer at the [Apache Software Foundation](http://www.apache.org/) and Open Source enthusiast.   
Most of what I write here is about Java based software development.   
  
The opinions expressed here are entirely my own and do not reflect the opinion of my employer or that of the Apache Software Foundation.

[View my complete profile](https://plus.google.com/101111641807482638731)

There was an error in this gadget

**ARCHIVE**

* [►](javascript:void(0))[2013](http://blog.jeroenreijn.com/search?updated-min=2013-01-01T00:00:00%2B01:00&updated-max=2014-01-01T00:00:00%2B01:00&max-results=1) (1)
* [►](javascript:void(0))[2012](http://blog.jeroenreijn.com/search?updated-min=2012-01-01T00:00:00%2B01:00&updated-max=2013-01-01T00:00:00%2B01:00&max-results=4) (4)
* [►](javascript:void(0))[2011](http://blog.jeroenreijn.com/search?updated-min=2011-01-01T00:00:00%2B01:00&updated-max=2012-01-01T00:00:00%2B01:00&max-results=8) (8)
* [▼](javascript:void(0))[2010](http://blog.jeroenreijn.com/search?updated-min=2010-01-01T00:00:00%2B01:00&updated-max=2011-01-01T00:00:00%2B01:00&max-results=9) (9)
  + [►](javascript:void(0))[October](http://blog.jeroenreijn.com/2010_10_01_archive.html) (3)
  + [►](javascript:void(0))[June](http://blog.jeroenreijn.com/2010_06_01_archive.html) (1)
  + [►](javascript:void(0))[May](http://blog.jeroenreijn.com/2010_05_01_archive.html) (2)
  + [▼](javascript:void(0))[April](http://blog.jeroenreijn.com/2010_04_01_archive.html) (1)
    - [Metadata extraction with Apache Tika](http://blog.jeroenreijn.com/2010/04/metadata-extraction-with-apache-tika.html)
  + [►](javascript:void(0))[February](http://blog.jeroenreijn.com/2010_02_01_archive.html) (1)
  + [►](javascript:void(0))[January](http://blog.jeroenreijn.com/2010_01_01_archive.html) (1)
* [►](javascript:void(0))[2009](http://blog.jeroenreijn.com/search?updated-min=2009-01-01T00:00:00%2B01:00&updated-max=2010-01-01T00:00:00%2B01:00&max-results=12) (12)
* [►](javascript:void(0))[2008](http://blog.jeroenreijn.com/search?updated-min=2008-01-01T00:00:00%2B01:00&updated-max=2009-01-01T00:00:00%2B01:00&max-results=24) (24)
* [►](javascript:void(0))[2007](http://blog.jeroenreijn.com/search?updated-min=2007-01-01T00:00:00%2B01:00&updated-max=2008-01-01T00:00:00%2B01:00&max-results=50) (51)
* [►](javascript:void(0))[2006](http://blog.jeroenreijn.com/search?updated-min=2006-01-01T00:00:00%2B01:00&updated-max=2007-01-01T00:00:00%2B01:00&max-results=50) (68)

**SUBSCRIBE TO**

http://img2.blogblog.com/img/widgets/arrow_dropdown.gifhttp://img1.blogblog.com/img/icon_feed12.png Posts

http://img2.blogblog.com/img/widgets/arrow_dropdown.gifhttp://img1.blogblog.com/img/icon_feed12.png Comments

© 2012 Jeroen Reijn. Powered by [Blogger](http://www.blogger.com/" \t "_blank).