Text and Metadata Extraction with Apache Tika

Jukka Zitting
Day Software

Background

● Day

Senior Developer

lucid
IMAGINATION

Technical Advisor

The Midgard Project

Apache Jackrabbit

Apache Tika

JPE
ZIP
PDF
Adobe
.MP3

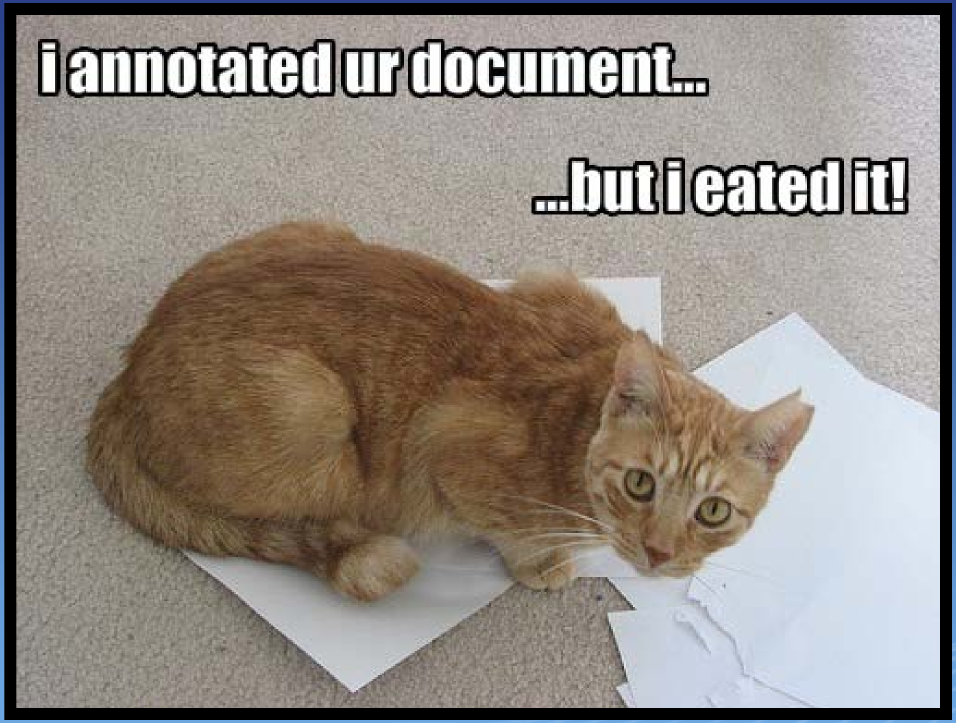
from files…

YOU don't know about me without you have read a book by the name of The Adventures of Tom Sawyer; but that ain't no matter. That book was made by Mr. Mark Twain, and he told the truth, mainly. There was things which he stretched, but mainly he told the truth. That is nothing. I never seen anybody but lied one time or another, without it was Aunt Polly, or the widow, or maybe Mary. Aunt Polly--Tom's Aunt Polly, she is--and Mary, and the Widow Douglas is all told about in that book, which is mostly a true book, with some stretchers, as I said before…

3 May. Bistritz.--Left Munich at 8:35 P.M., on 1st May, arriving at Vienna early next morning; should have arrived at 6:46, but train was an hour late. Buda-Pesth seems a wonderful place, from the glimpse which I got of it from the train and the little I could walk through the streets. I feared to go very far from the station, as we had arrived late and would start as near the correct time as possible. The impression I had was that we were leaving the West and entering the East; the most western of splendid bridges over the Danube, which is here of noble width and depth, took us among the traditions of Turkish rule…

It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife. However little known the feelings or views of such a man may be on his first entering a neighbourhood, this truth is so well fixed in the minds of the surrounding families, that he is considered the rightful property of some one or other of their daughters. "My dear Mr. Bennet," said his lady to him one day, "have you heard that Netherfield Park is let at last?" Mr. Bennet replied that he had not. "But it is," returned she; "for Mrs. Long has just been here, and she told me all about it." Mr. Bennet made no answer…

… to text …



© Doug Schepers

… and metadata

<Apache Tika/>

The Agenda

Tika in a nutshell

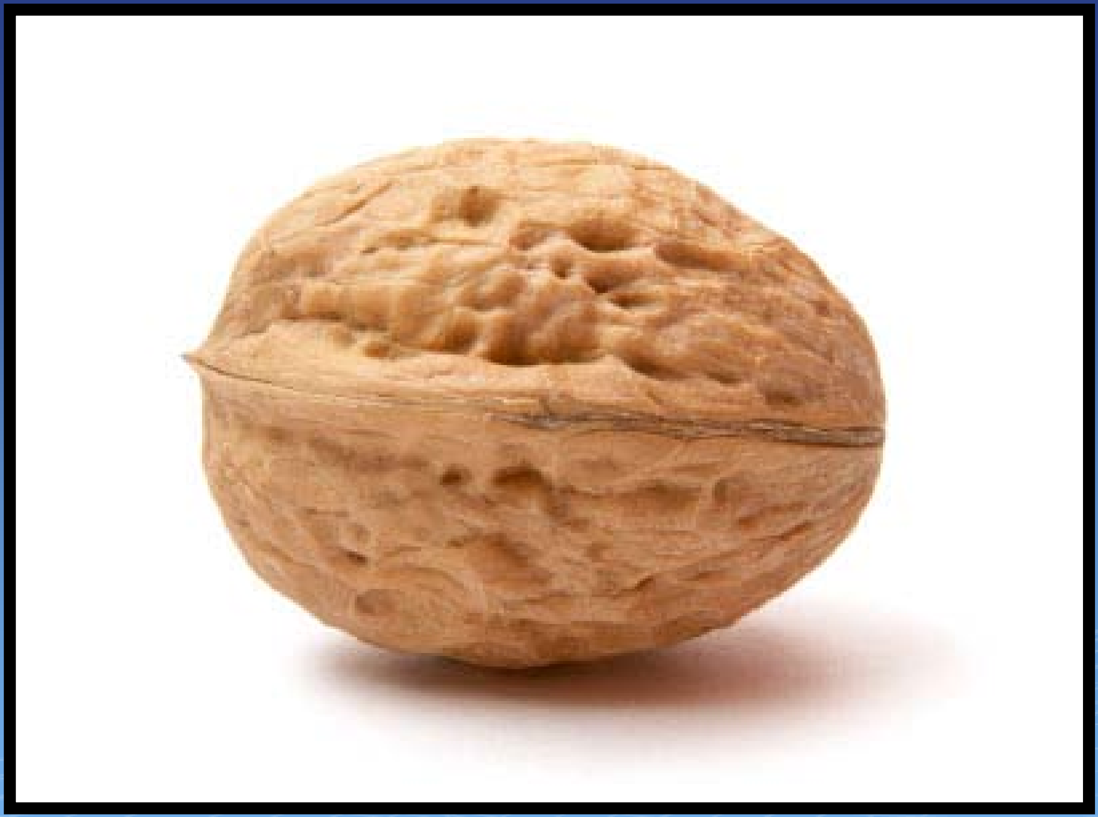
Tika in action:

Command line and GUI -

The Tika façade -

The Parser API –

Solr Cell -



Tika in a nutshell

Some History

-2006 Initial discussions about Tika

2007 Project starts in the Apache Incubator

2008 Releases 0.1 and 0.2, graduates into a Lucene subproject

2009 Releases 0.3, 0.4 and 0.5

2010 (so far) Releases 0.6 and 0.7, becomes an Apache TLP

Some Statistics

8 committers, 101 contributors (more welcome!)

17kLOC + 10k lines of comments, written in 708 commits

250 classes in 32 packages, 60% test coverage

3 mailing lists, ~150 msgs per month (dev 100, use 30, svn 20)

1277 known media types + 51 aliases

942 filename globs, 310 magic byte patterns, 16 known XML root elements

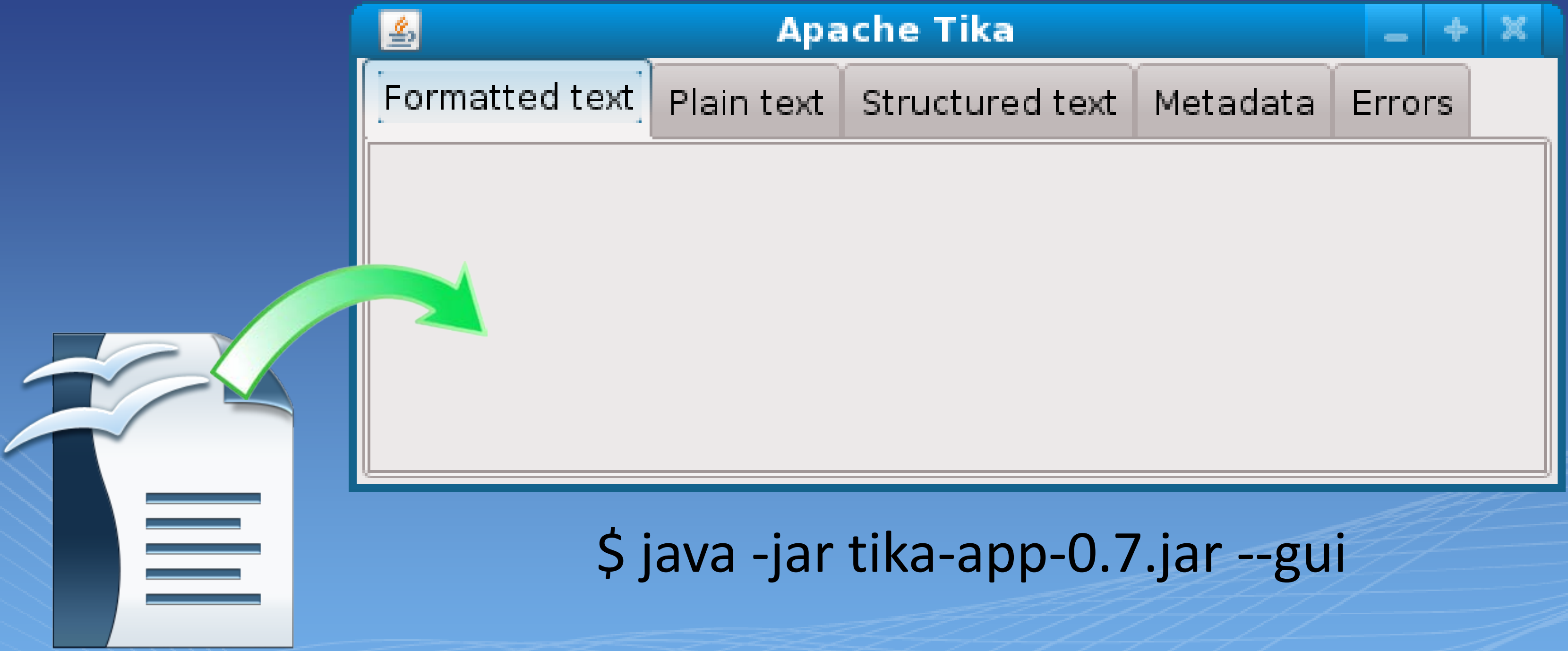
parser support for all major document formats (and many more)

THE FUTURE
BELONGS TO
THE FEW OF US
STILL WILLING
TO GET OUR
HANDS DIRTY.

tika-app-0.7.jar (17MB)

```
$ java -jar tika-app-0.7.jar --xhtml /path/to/document.doc

$ java -jar tika-app-0.7.jar --text http://example.org/doc

$ java -jar tika-app-0.7.jar --metadata < document.doc

$ cat document.doc | java -jar tika-app-0.7.jar --text | grep foo

$ java -jar tika-app-0.7.jar --help
```

Tika on the command line



```
$ java -jar tika-app-0.7.jar --gui
```

Tika GUI

Tika façade

```
import org.apache.tika.Tika;

Tika tika = new Tika();

String type = tika.detect(…);

Reader reader = tika.parse(…);

String text = tika.parseToString(…);
```

Where … can be:

java.lang.String

java.io.File

java.net.URL

java.io.InputStream

Dependency management

- tika-app-0.7.jar – simple and easy
- For more control, use Maven or Ivy

```
<dependency>
  <groupId>org.apache.tika</groupId>
  <artifactId>tika-parsers</artifactId>
  <version>0.7</version>
</depenency>
```

- Comes with log4j, etc.

Dependencies listed

tika-core-0.7.jar
tika-parsers-0.7.jar
tagsoup-1.2.jar
asm-3.1.jar
xmlbeans-2.3.0.jar
dom4j-1.6.1.jar
xml-apis-1.0.b2.jar
log4j-1.2.14.jar

pdfbox-1.1.0.jar
fontbox-1.1.0.jar
jempbox-1.1.0.jar
bcmail-jdk15-1.45.jar
bcprov-jdk15-1.45.jar
poi-3.6.jar
poi-scratchpad-3.6.jar
poi-ooxml-3.6.jar

poi-ooxml-schemas-3.6.jar
commons-compress-1.0.jar
metadata-extractor-2.4.0-beta-1.jar
geronimo-stax-api_1.0_spec-1.0.1.jar
commons-logging-1.1.1.jar

… yes, that’s 21 jars

```
java.io.InputStream input = …;

org.xml.sax.ContentHandler handler = …;

Metadata metadata = new Metadata();

ParseContext context = new ParseContext();

Parser parser = new AutoDetectParser();

parser.parse(input, handler, metadata, context);
```

Tika Parser API

```
java.io.InputStream input = …;

org.xml.sax.ContentHandler hand

Metadata metadata = new Metadata();

ParseContext context = new ParseContext();

Parser parser = new AutoDetectParser();

parser.parse(input, handler, metadata, context);
```

The input document.
Note: The stream won’t be closed!

Tika Parser API

```
import org.apache.tika.io.TikaInputStream;

InputStream input = new TikaInputStream(…);
```

- For parsers that need the whole file
- Automatic input metadata

Where … can be:

java.lang.String

java.io.File

java.net.URL

java.io.InputStream

TikaInputStream ⚠ new in Tika 0.8

Tika Parser API

```
java.io.InputStream input = …;
org.xml.sax.ContentHandler handler = …;
Metadata metadata = new Metad
ParseContext context = new ParseContext();
Parser parser = new AutoDetectParser();
parser.parse(input, handler, metadata, context);
```

XHTML SAX event handler
for the extracted structured text.

```
<html xmlns="http://www.w3.org/1999/xhtml">

<head><title>…</title></head>

<body>…</body>

</html>
```

- SAX = streaming support
- XHTML = structured, semantic
- Not designed for rendering!
- Not 1-to-1 with input HTML!

XHTML SAX events

```
java.io.InputStream input = …;

org.xml.sax.ContentHandler handler = …;

Metadata metadata = new Metadata();

ParseContext context = new Parse

Parser parser = new AutoDetectParser();

parser.parse(input, handler, metadata, context);
```

Document metadata, both for input (filename) and output (title)

Tika Parser API

```
import org.apache.tika.metadata.Metadata;

Metadata metadata = new Metadata();

metadata.set(Metadata.RESOURCE_NAME_KEY, "…");

String title = metadata.get(Metadata.TITLE);

String type = metadata.get(Metadata.CONTENT_TYPE);
```

Tika Metadata API

Metadata

- Media type (text/plain, application/pdf, etc.)
- Title, Author, Subject, Date, Copyright, etc. (Dublin Core)
- Photos/Images: Size, Depth, Color space, Camera settings, etc. (EXIF)
- Video/Audio: Frame rate, Duration, Codec, etc.
- XMP?

Tika Parser API

```
java.io.InputStream input = …;
org.xml.sax.ContentHandler handler = …;
Metadata metadata = new Metadata();
ParseContext context = new ParseContext();
Parser parser = new AutoDetectPa
parser.parse(input, handler, metadata, context);
```

Parsing context for extra options passed to the parser instances

```
import org.apache.tika.parser.ParseContext;

ParseContext context = new ParseContext();

context.set(HtmlMapper.class, new MyHtmlMapper());

context.set(Parser.class, new MyParser());

context.set(Locale.class, Locale.CZ);
```

Using the parse context

Tika Parser API

```
java.io.InputStream input = …;
org.xml.sax.ContentHandler handler = …;
Metadata metadata = new Metadata();
ParseContext context = new ParseContext();
Parser parser = new AutoDetectParser();
parser.parse(input, handler, meta
```

Automatically selects best parser based on detected document type

Parser libraries (AL-compatible)

PDF – Apache PDFBox

MS Office – Apache POI

HTML – Tagsoup

Images – ImageIO, metadata-extractor

Zip, Tar, Gz, etc. – Commons Compress

etc.

```
java.io.InputStream input = …;

org.xml.sax.ContentHandler handler = …;

Metadata metadata = new Metadata();

ParseContext context = new ParseContext();

Parser parser = new AutoDetectParser();

parser.parse(input, handler, metadata, context);
```

throws TikaException, IOException, SAXException

Tika Parser API

Apache Solr

```
$ curl http://localhost:8983/solr/update/extract?literal.id=doc1 \
          -F file=@document.doc
```

Solr Cell (extracting request handler)

Apache Nutch

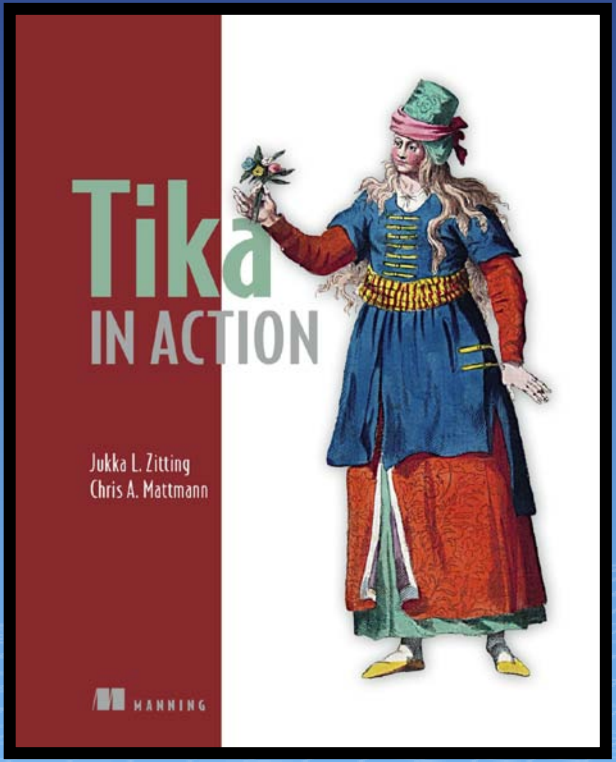
Apache Jackrabbit

Apache UIMA

etc.

Other Integrations

Questions?



Tika
IN ACTION

Jukka L. Zitting
Chris A. Mattmann

MANNING

MEAP starting soon!

Extras

- 1-1 mapping of input HTML
- Parsing documents from inside packages

```
import org.apache.tika.parser.html.*;

ParseContext context = new ParseContext();

context.set(HtmlMapper.class, IdentityHtmlMapper.INSTANCE);



Parser parser = …;

parser.parse(…, …, …, context);
```

1-1 mapping of input HTML

```
ParseContext context = new ParseContext();

context.set(Parser.class, new MyComponentParser());



Parser parser = …;

parser.parse(…, …, …, context);
```

Parsing documents from inside packages