The secrets of a file

Jukka Zitting | contributor to Tika, PDFBox, POI, Commons, TagSoup, junrar, etc.

Agenda

- Type detection
- Metadata extraction
- Future work: XMP

Type detection

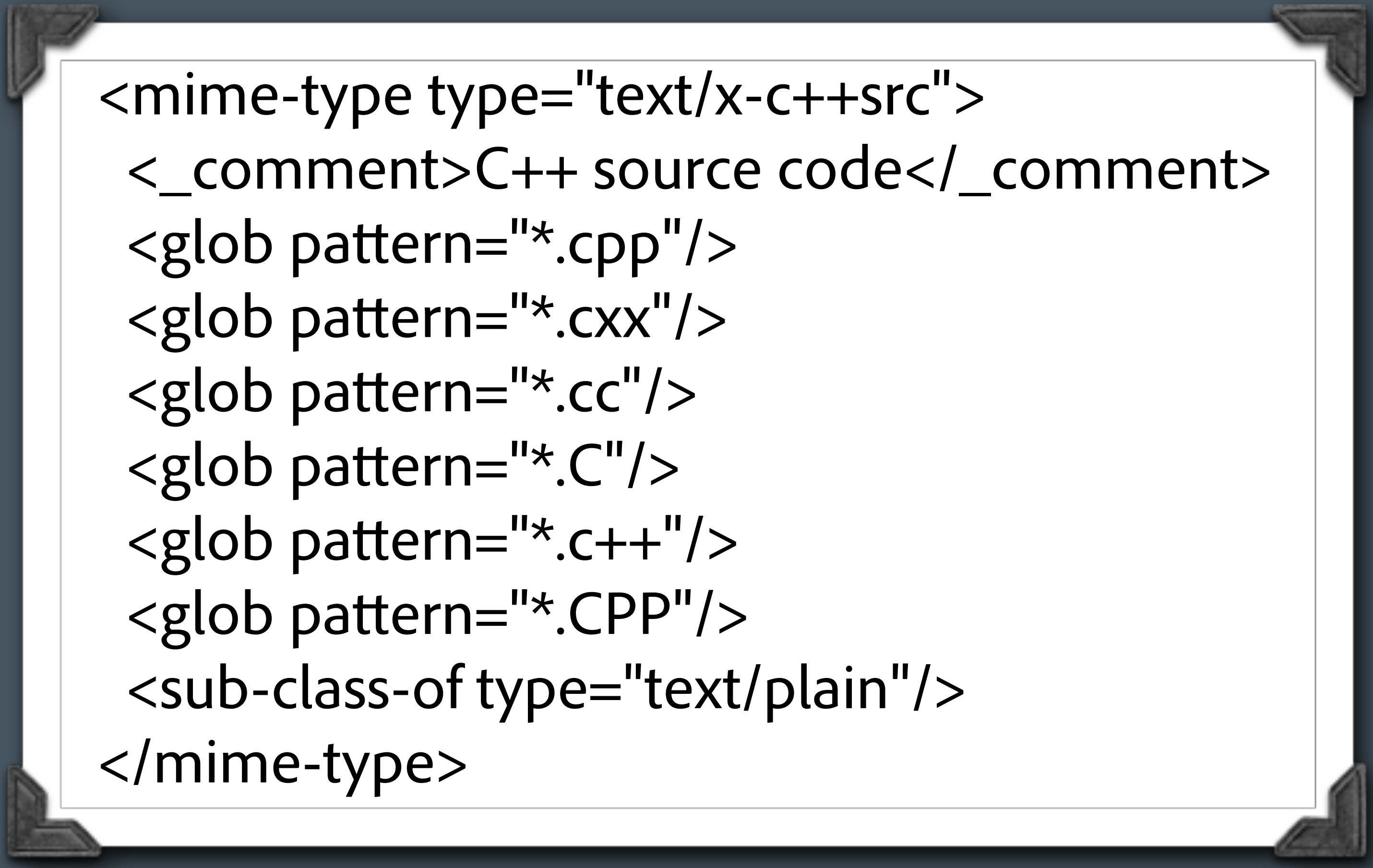
Demo

File extensions

Source code: one standard extension

```
<mime-type type="text/x-java-source">
  <_comment>Java source code</_comment>
  <alias type="text/x-java" />
  <glob pattern="*.java"/>
  <sub-class-of type="text/plain"/>
 </mime-type>
```

Source code: multiple commonly used extensions



```
<mime-type type="text/x-c++src">
  <_comment>C++ source code</_comment>
  <glob pattern="*.cpp"/>
  <glob pattern="*.cxx"/>
  <glob pattern="*.cc"/>
  <glob pattern="*.C"/>
  <glob pattern="*.c++"/>
  <glob pattern="*.CPP"/>
  <sub-class-of type="text/plain"/>
</mime-type>
```

Source code: conflicting extensions

```
 <mime-type type="text/x-prolog">
  <_comment>Prolog source</_comment>
  <glob pattern="*.pro"/>
  <!-- <glob pattern="*.pl"/>
       - conflicts with text/x-perl -->
  <sub-class-of type="text/plain"/>
</mime-type>
```

Common first step also for other types of files

```
<mime-type type="image/x-raw-canon">
  <_comment>Canon raw image</_comment>
  <glob pattern="*.crw"/>
  <glob pattern="*.cr2"/>
</mime-type>
```

No file name?
Use magic!

One standard byte pattern

```
<mime-type type="application/pdf">
  <alias type="application/x-pdf"/>
  <acronym>PDF</acronym>
  <_comment>Portable Doc...</_comment>
  <magic priority="50">
    <match value="%PDF-"
                   type="string" offset="0"/>
  </magic>
  <glob pattern="*.pdf"/>
</mime-type>
```

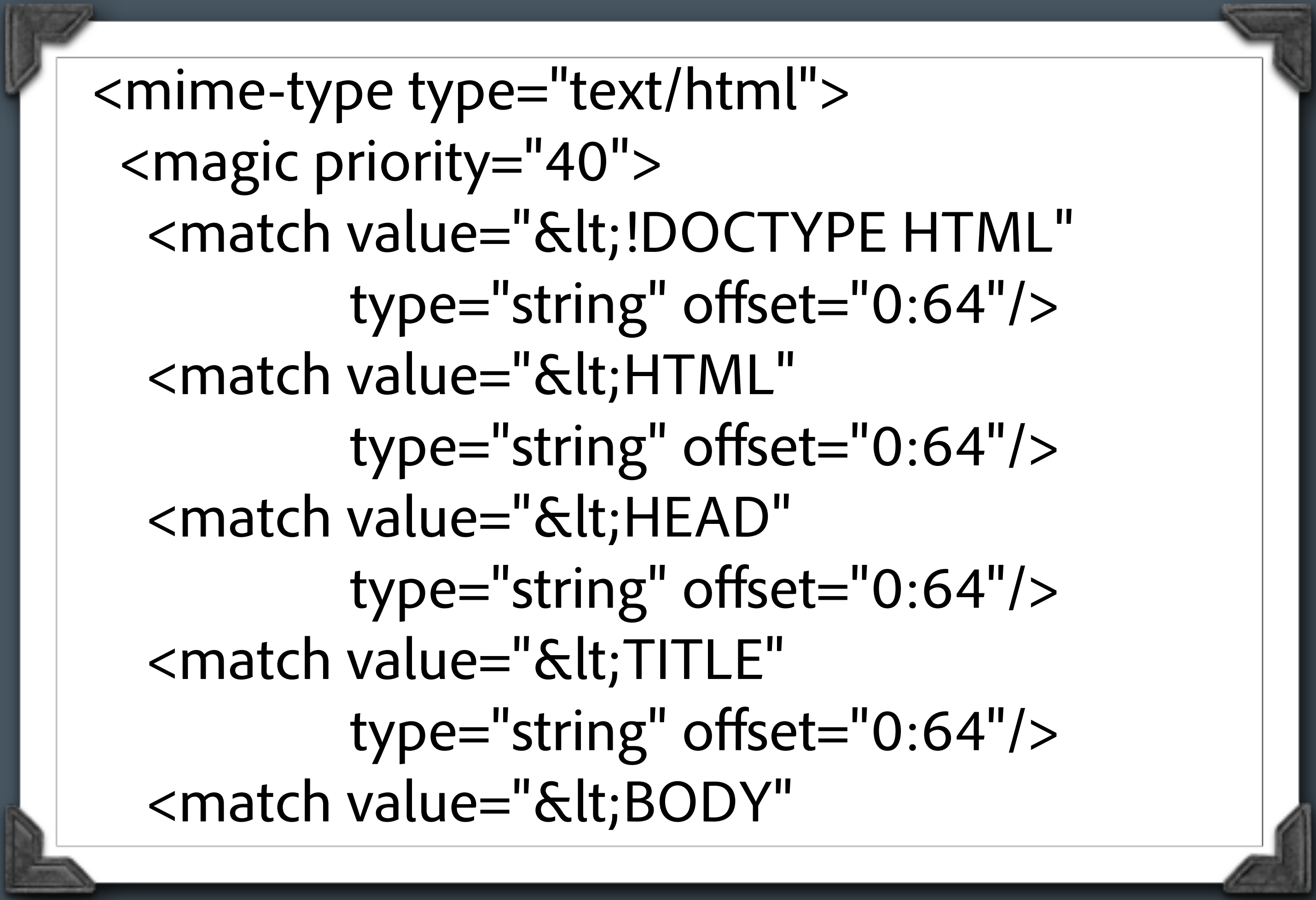
Alternative byte patterns

```
<mime-type type="image/gif">
 <acronym>GIF</acronym>
 <_comment>Graphics Inter...</_comment>
 <magic priority="50">
   <match value="GIF87a"
                type="string" offset="0"/>
   <match value="GIF89a"
                type="string" offset="0"/>
 </magic>
 <glob pattern="*.gif"/>
</mime-type>
```

Odd cases: MS Word

```
<mime-type type="application/msword">
  <alias type="application/vnd.ms-word"/>
  <magic priority="50">
    <match value="Microsoft\ Word\ 6.0..."
                 type="string" offset="2080"/>
    <match value="Documento\ Microsoft..."
                 type="string" offset="2080"/>
    <match value="MSWordDoc"
                 type="string" offset="2112"/>
    <match value="0x31be0000"
                 type="big32" offset="0"/>
```

Odd cases: HTML



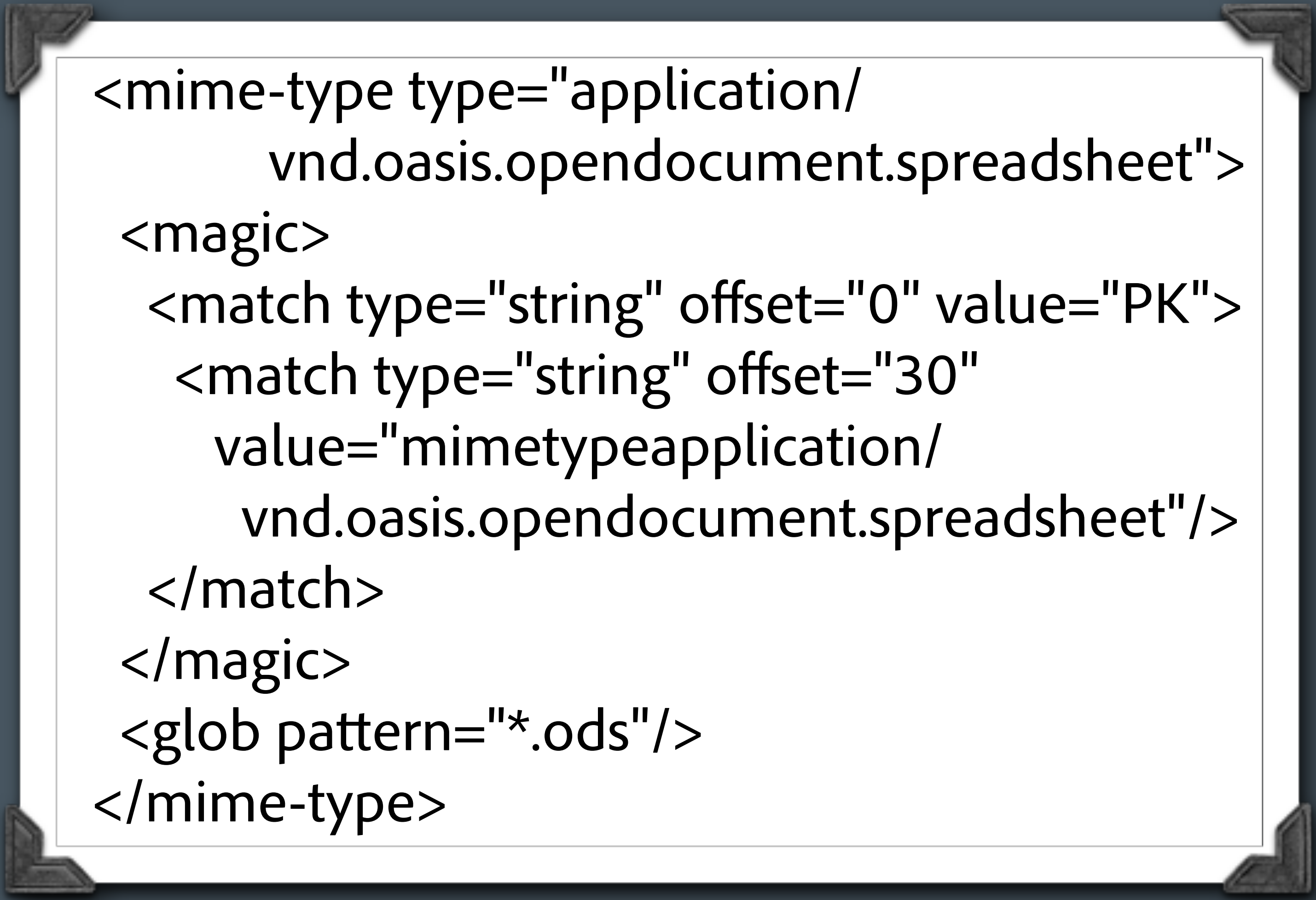
```
<mime-type type="text/html">
  <magic priority="40">
    <match value="&lt;!DOCTYPE HTML"
                   type="string" offset="0:64"/>
    <match value="&lt;HTML"
                   type="string" offset="0:64"/>
    <match value="&lt;HEAD"
                   type="string" offset="0:64"/>
    <match value="&lt;TITLE"
                   type="string" offset="0:64"/>
    <match value="&lt;BODY"
```

Container formats

XML formats

```
<mime-type type="application/xhtml+xml">
  <magic priority="50">
    <match value="&lt;html xmlns="
                  type="string" offset="0:8192"/>
  </magic>
  <root-XML namespaceURI=
                  "http://www.w3.org/1999/xhtml"
                  localName="html"/>
  <glob pattern="*.xhtml"/>
  <glob pattern="*.xht"/>
</mime-type>
```

ZIP archives



```
<mime-type type="application/
           vnd.oasis.opendocument.spreadsheet">
  <magic>
    <match type="string" offset="0" value="PK">
      <match type="string" offset="30"
         value="mimetypeapplication/
           vnd.oasis.opendocument.spreadsheet"/>
    </match>
  </magic>
  <glob pattern="*.ods"/>
</mime-type>
```

Custom cases

Custom cases

- Zip archives
  - Parse the container and look at contained file names
- Old MS Office formats
  - Parse the container and look at contained resources
- Plain text
  - Really tricky, see below...

Detecting plain text

- UTF BOMs
- Control characters
- Line endings
- Byte histogram
- Not foolproof, but quite reliable in practice

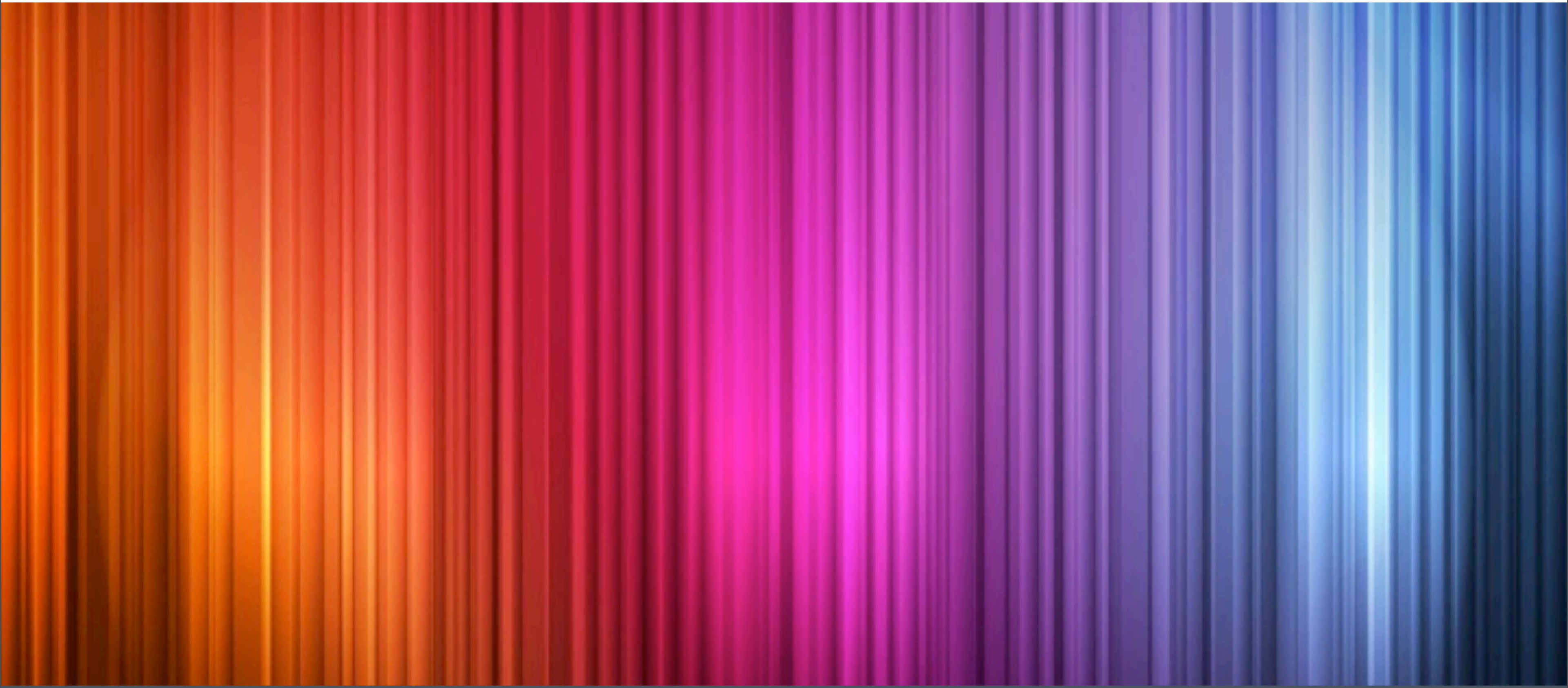
Composite approach

The ultimate type detector

- Custom cases
- Magic byte patterns
- Container types
- File name hints
- Content type hints
- application/octet-stream

- Less reliable detectors can only add detail to more reliable results

Metadata extraction



Demo

Types of metadata

Dublin core

- Basic information
  - dc:title
  - dc:creator
  - dc:date
  - dc:format
- Driven originally by (scientific) libraries
- http://dublincore.org/

Exif

- Image metadata
  - tiff:ImageLength
  - tiff:ImageWidth
  - tiff:BitsPerSample
  - ...
- Useful also for non-TIFF/JPEG image formats

Other schemas

- International Press Telecommunications Council (IPTC)
- Various XMP-related schemas

Future work

XMP

Adobe