Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY OF NEW ZEALAND

Metadata Extraction Tool Changes

Version*: 3.5.*

Metadata Extraction Tool Changes

Version: 3.5.

Table of Contents

✓ *Introduction*

This document covers the changes to the Metadata Extraction Tool.

✓ *Version 3.5*

Version 3.5 has the following new features added into the Metadata Extractor Tool

- Addition of new Extractors for the following file formats
  FLAC, Arc and MP3 ID3Lyric
- Enhancements to the existing Wav extractor to be able to extract metadata from Broadcast Wave Format
- Minor defect fixes

✓ *Version 3.4*

Version 3.4 corrects several MIME types reported for Microsoft Office products.

✓ *Version 3.3*

Version 3.3 of the Metadata Extraction Tool adds the following:

- Corrects a number of adapters that were not correctly closing files. This could cause issues when the tool was embedded in other applications.
- A set of unit tests that can be run with JUnit 4.

✓ *Version 3.2*

Version 3.2 makes the following enhancement on v3.1.

- Minor modification to allow the configuration file to be read from directories containing spaces.

✓ *Version 3.1*

Version 3.1 makes the following enhancements on v3.0.

File Support

- The Metadata Extraction Tool defaults to using an open source PDF Parser named PDFBox (http://www.pdfbox.org/). This library resolves issues with being unable to parse encrypted PDFs and has an improved reliability over the old parsing mechanism.
- The JPG adapter XSLT has been altered to remove extraneous namespace information.
- The batch scripts for running the Metadata Tool are now able to run even under directories including spaces.
- The TIFF adapter no longer extracts the StripByteCounts elements. These elements are data required to read the image, but don't provide information about the image itself.

✓ *Version 3.0*

Version 3.0 is the initial release to the Open Source community.

File Support

Version 3.0 of the Metadata Extraction Tool changes the way the files are identified by the tool. In earlier versions, the Metadata Extractor tool relied heavily on

extensions. The new version relies on magic numbers where possible. Notable exceptions to this rule are:

- MP3 files are still detected by their extension as using the magic number (0xFF at the start of the file) risks a large number of false positives.
- Microsoft Office and Works documents are identified based on their extension, plus a check to ensure that they are valid OLE files. This is due to difficulties in being able to consistently locate metadata that identifies the type of the document. While this information is available, it varies in its position.
- There have been major changes to the XMLAdapter. The new version will extract the information from the XML and DOCTYPE declarations, while previous versions essentially copied the source XML file.

Major Changes

Version 3.0 includes the following major changes:

- Additional JavaDocs, especially in the interfaces where development is expected.
- New nz.govt.natlib.samples package containing usage samples.
- Update and addition of new documentation.
- Changes to the XMLAdapter to extract the XML version and encoding types and extract information from the DOCTYPE header.
- Creation of ANT build scripts to automate the build process.
- Modifications to the default build to include all adapters.