TIKA-1234

TWLT

2014-02-10

Contents

Chapter 1

Root issue TIKA-1234

1.1 Summary

empty docx creates exception

1.2 Description

using an empty docx File as input results in exception. Trace:

Apache Tika was unable to parse the document
at F:\Microsoft Word-Dokument (neu) (2).docx.

The full exception stack trace is included below:

```
org.apache.tika.exception.TikaException: Error creating OOXML extractor
at org.apache.tika.parser.microsoft.ooxml.OOXMLExtractorFactory.parse(OOXMLExtractorFactory.java:128)
at org.apache.tika.parser.microsoft.ooxml.OOXMLParser.parse(OOXMLParser.java:82)
at org.apache.tika.parser.CompositeParser.parse(CompositeParser.java:242)
at org.apache.tika.parser.CompositeParser.parse(CompositeParser.java:242)
at org.apache.tika.parser.AutoDetectParser.parse(AutoDetectParser.java:120)
at org.apache.tika.gui.TikaGUI.handleStream(TikaGUI.java:320)
at org.apache.tika.gui.TikaGUI.openFile(TikaGUI.java:279)
at org.apache.tika.gui.ParsingTransferHandler.importFiles(ParsingTransferHandler.java:94)
at org.apache.tika.gui.ParsingTransferHandler.importData(ParsingTransferHandler.java:77)
at javax.swing.TransferHandler.importData(Unknown Source)
at javax.swing.TransferHandler$DropHandler.drop(Unknown Source)
at java.awt.dnd.DropTarget.drop(Unknown Source)
at javax.swing.TransferHandler$SwingDropTarget.drop(Unknown Source)
at sun.awt.dnd.SunDropTargetContextPeer.processDropMessage(Unknown Source)
at sun.awt.dnd.SunDropTargetContextPeer$EventDispatcher.dispatchDropEvent(Unknown Source)
at sun.awt.dnd.SunDropTargetContextPeer$EventDispatcher.dispatchEvent(Unknown Source)
at sun.awt.dnd.SunDropTargetEvent.dispatch(Unknown Source)
at java.awt.Component.dispatchEventImpl(Unknown Source)
at java.awt.Container.dispatchEventImpl(Unknown Source)
at java.awt.Component.dispatchEvent(Unknown Source)
at java.awt.LightweightDispatcher.retargetMouseEvent(Unknown Source)
at java.awt.LightweightDispatcher.processDropTargetEvent(Unknown Source)
at java.awt.LightweightDispatcher.dispatchEvent(Unknown Source)
at java.awt.Container.dispatchEventImpl(Unknown Source)
at java.awt.Window.dispatchEventImpl(Unknown Source)
at java.awt.Component.dispatchEvent(Unknown Source)
at java.awt.EventQueue.dispatchEventImpl(Unknown Source)
at java.awt.EventQueue.access$200(Unknown Source)
at java.awt.EventQueue$3.run(Unknown Source)
at java.awt.EventQueue$3.run(Unknown Source)
at java.security.AccessController.doPrivileged(Native Method)
at java.security.ProtectionDomain$1.doIntersectionPrivilege(Unknown Source)
at java.security.ProtectionDomain$1.doIntersectionPrivilege(Unknown Source)
at java.awt.EventQueue$4.run(Unknown Source)
at java.awt.EventQueue$4.run(Unknown Source)
at java.security.AccessController.doPrivileged(Native Method)
at java.security.ProtectionDomain$1.doIntersectionPrivilege(Unknown Source)
at java.awt.EventQueue.dispatchEvent(Unknown Source)
at java.awt.EventDispatchThread.pumpOneEventForFilters(Unknown Source)
at java.awt.EventDispatchThread.pumpEventsForFilter(Unknown Source)
at java.awt.EventDispatchThread.pumpEventsForHierarchy(Unknown Source)
at java.awt.EventDispatchThread.pumpEvents(Unknown Source)
at java.awt.EventDispatchThread.pumpEvents(Unknown Source)
at java.awt.EventDispatchThread.run(Unknown Source)
Caused by: org.apache.poi.openxml4j.exceptions.InvalidFormatException: Package should contain a content type part [M1.13]
at org.apache.poi.openxml4j.opc.ZipPackage.getPartsImpl(ZipPackage.java:178)
at org.apache.poi.openxml4j.opc.OPCPackage.getParts(OPCPackage.java:662)
at org.apache.poi.openxml4j.opc.OPCPackage.open(OPCPackage.java:269)
at org.apache.tika.parser.microsoft.ooxml.OOXMLExtractorFactory.parse(OOXMLExtractorFactory.java:74)
... 43 more
```

1.3 Commits

No related commits

1.4 Comments

1. **nick:** What do you mean by an "empty docx" file? Fire up word, new document, then save before adding any content? Or creating a 0 byte file and calling it foo.docx?

2. **TWLT:** I created a new docx in Windows through the context menu New->Microsoft Word Document.
   It seems it has something to do with that because I couldn't recreate it with an empty file created directly in Word.

3. **nick:** Is that a 0 byte file?

   If so, that isn't a valid .docx document, so it's expected that Tika will break on it. IIRC, the minimum size for a .doc file is about 3kb (something like 6 x 512 byte ole2 frames), and for a .docx will be something similar (for the compressed minimal xml files)

4. **TWLT:** you're right it's 0 byte. My mistake, didn't think Windows is created a wrong docx there.

1.5 Pull requests

No pull requests

Chapter 2

Connected issue TIKA-1235

2.1 Summary

empty docx creates exception

2.2 Description

using an empty docx File as input results in exception. Trace:

```
Apache Tika was unable to parse the document
at F:\Microsoft Word-Dokument (neu) (2).docx.

The full exception stack trace is included below:

org.apache.tika.exception.TikaException: Error creating OOXML extractor
at org.apache.tika.parser.microsoft.ooxml.OOXMLExtractorFactory.parse(OOXMLExtractorFactory.java:128)
at org.apache.tika.parser.microsoft.ooxml.OOXMLParser.parse(OOXMLParser.java:82)
at org.apache.tika.parser.CompositeParser.parse(CompositeParser.java:242)
at org.apache.tika.parser.CompositeParser.parse(CompositeParser.java:242)
at org.apache.tika.parser.AutoDetectParser.parse(AutoDetectParser.java:120)
at org.apache.tika.gui.TikaGUI.handleStream(TikaGUI.java:320)
at org.apache.tika.gui.TikaGUI.openFile(TikaGUI.java:279)
at org.apache.tika.gui.ParsingTransferHandler.importFiles(ParsingTransferHandler.java:94)
at org.apache.tika.gui.ParsingTransferHandler.importData(ParsingTransferHandler.java:77)
at javax.swing.TransferHandler.importData(Unknown Source)
at javax.swing.TransferHandler$DropHandler.drop(Unknown Source)
at java.awt.dnd.DropTarget.drop(Unknown Source)
at javax.swing.TransferHandler$SwingDropTarget.drop(Unknown Source)
at sun.awt.dnd.SunDropTargetContextPeer.processDropMessage(Unknown Source)
at sun.awt.dnd.SunDropTargetContextPeer$EventDispatcher.dispatchDropEvent(Unknown Source)
at sun.awt.dnd.SunDropTargetContextPeer$EventDispatcher.dispatchEvent(Unknown Source)
at sun.awt.dnd.SunDropTargetEvent.dispatch(Unknown Source)
at java.awt.Component.dispatchEventImpl(Unknown Source)
at java.awt.Container.dispatchEventImpl(Unknown Source)
at java.awt.Component.dispatchEvent(Unknown Source)
at java.awt.LightweightDispatcher.retargetMouseEvent(Unknown Source)
at java.awt.LightweightDispatcher.processDropTargetEvent(Unknown Source)
at java.awt.LightweightDispatcher.dispatchEvent(Unknown Source)
at java.awt.Container.dispatchEventImpl(Unknown Source)
at java.awt.Window.dispatchEventImpl(Unknown Source)
at java.awt.Component.dispatchEvent(Unknown Source)
at java.awt.EventQueue.dispatchEventImpl(Unknown Source)
at java.awt.EventQueue.access$200(Unknown Source)
at java.awt.EventQueue$3.run(Unknown Source)
at java.awt.EventQueue$3.run(Unknown Source)
at java.security.AccessController.doPrivileged(Native Method)
at java.security.ProtectionDomain$1.doIntersectionPrivilege(Unknown Source)
at java.security.ProtectionDomain$1.doIntersectionPrivilege(Unknown Source)
at java.awt.EventQueue$4.run(Unknown Source)
at java.awt.EventQueue$4.run(Unknown Source)
at java.security.AccessController.doPrivileged(Native Method)
at java.security.ProtectionDomain$1.doIntersectionPrivilege(Unknown Source)
at java.awt.EventQueue.dispatchEvent(Unknown Source)
at java.awt.EventDispatchThread.pumpOneEventForFilters(Unknown Source)
at java.awt.EventDispatchThread.pumpEventsForFilter(Unknown Source)
at java.awt.EventDispatchThread.pumpEventsForHierarchy(Unknown Source)
at java.awt.EventDispatchThread.pumpEvents(Unknown Source)
at java.awt.EventDispatchThread.pumpEvents(Unknown Source)
at java.awt.EventDispatchThread.run(Unknown Source)
Caused by: org.apache.poi.openxml4j.exceptions.InvalidFormatException: Package should contain a content type part [M1.13]
at org.apache.poi.openxml4j.opc.ZipPackage.getPartsImpl(ZipPackage.java:178)
at org.apache.poi.openxml4j.opc.OPCPackage.getParts(OPCPackage.java:662)
at org.apache.poi.openxml4j.opc.OPCPackage.open(OPCPackage.java:269)
at org.apache.tika.parser.microsoft.ooxml.OOXMLExtractorFactory.parse(OOXMLExtractorFactory.java:74)
... 43 more
```

2.3 Commits

No related commits

2.4 Comments

1. **TWLT:** Sorry, can't provide the file. Upload is rejected with:
   Cannot attach empty file emtpy.docx.

2. **tpalsulich:** Closing as Invalid, since the file is 0 bytes (minimum, according to [~nick], is about 3kb).

2.5 Pull requests

No pull requests