OOXML Document Analysis

By Gülşah Köse

Software Engineer
@gulsahkse
How do we proceed when we receive a problematic OOXML document?
Reproduce the problem

There may be more than one problem in same document. We must make sure we produce the right one.
Simplify the document

The document may contain texts, images, sounds. But the problem might just about the images. If possible, other elements in the document can be cleaned.
Read the XML

OOXML documents (docx, pptx, xlsx etc) contains compressed XML files. Result of the “unzip test.docx” command will be seen like this.
Detect the XML part problematic element

For example: A paragraph element will look like as following

```xml
<w:p>
  <w:pPr>
    <w:pStyle w:val="Normal"/>
    <w:bidi w:val="0"/>
    <w:jc w:val="left"/>
    <w:rPr/>
  </w:pPr>
  <w:r>
    <w:rPr/>
    <w:t>Hello</w:t>
  </w:r>
</w:p>
```
Read the OOXML specs

It can be understood at this stage whether the syntax is appropriate or not. Apart from that, it allows us to understand the element in more detail than what I see in the interface.
Pick the keywords

We will need this to understand how Collabora Office handles these elements.
grep, read, debug loop

It cannot be formulated after this stage.

Reading code → detecting suspicious parts → debug them is a loop.

Often writing code is a small part of that process :)

(text/date via Insert>Header and Footer)
Real example

Bug report: https://bugs.documentfoundation.org/show_bug.cgi?id=142716
(PPTX EXPORT: Text content is partially lost)
Before

xxx and yyy

After

xxx and
xxx and
  Entrance: Dissolve in
<p:animEffect transition="in" filter="dissolve">
  <p:cBhvr>
    <p:cTn id="7" dur="500"/>
    <p:tgtEl>
      <p:spTgt spid="3">
        <p:txEl>
          <p:pRg st="0" end="0"/>
        </p:txEl>
        <p:spTgt>
      </p:tgtEl>
    </p:spTgt>
  </p:cBhvr>
</p:animEffect>
Solution:  https://gerrit.libreoffice.org/c/core/+/117557
Thanks!

By Gülşah Köse

@CollaboraOffice
hello@collaboraoffice.com
Collaboraoffice.com