

GSoC project: Improving RTF Import Presentation of a LibreOffice student

Miklos Vajna

13 October 2011

- Background
- Activities in LibreOffice earlier
- Project this summer: development of a new RTF import filter
 - developer side
 - user side

- I'm a student from Budapest University of Technology and Economics, Hungary
- A few project I am interested in:
 - LibreOffice – packaging, RTF filters
 - swig – a binding generator
 - git – I developed the current `git merge`
 - BitlBee – an IM ↔ IRC gateway
 - Frugalware Linux – a distribution

- RTF export filter in Writer
- Git-related patches
- Packager for Frugalware Linux

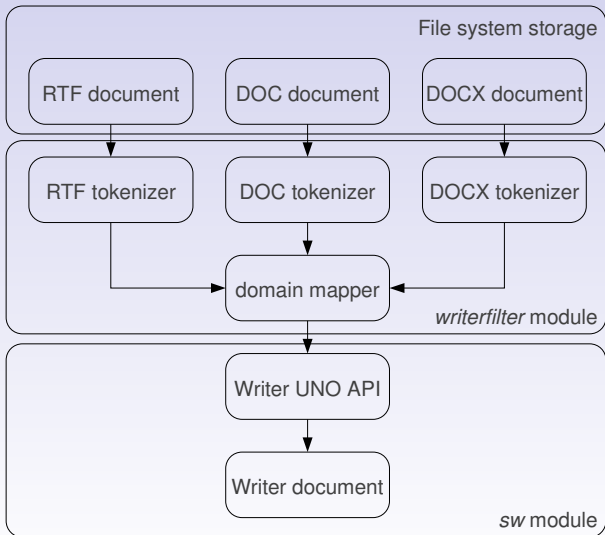
RTF Import Development

Summary

- The idea: RTF export is already subclassed from a generic Word exporter, the same could be done for the import
- The `writerfilter` module already provides `dmapper` for common Word vs. Writer problems (e.g. field parsing)
- Goal: support everything which was provided by the old filter, smaller size, new features

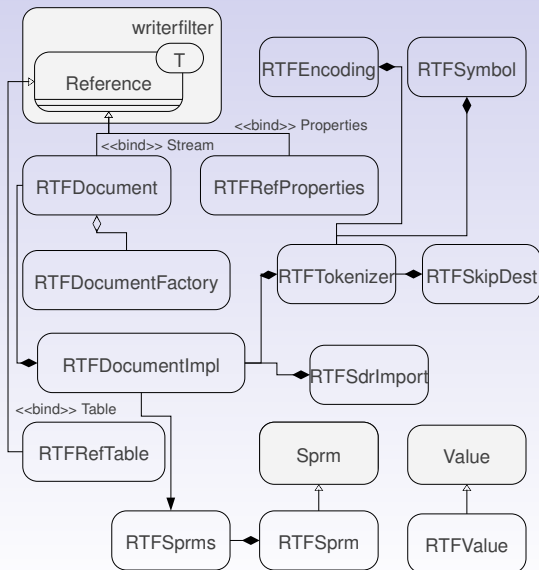
RTF Import Development

The big picture



RTF Import Development

Classes of the RTF import



- Created a unit test: it can quickly test if the tokenizer handles a document or not
- Can be run without building the `SW` module, even
- Does not replace manual testing (if the result visually matches the original)
- Documents produced by OpenOffice.org 3.3, LibreOffice 3.4, Word 2007, Word 2010

RTF Import New Features

Nested tables

Before:

Nested-table:¶	
A¶	
B¶	
¶	
C¶	
D¶	
¶	
C¶	B'D¶ ¶

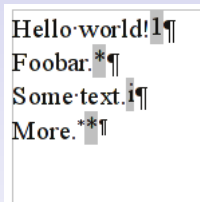
After:

Nested-table:¶		
A¶	B¶	B¶
C¶	D¶	
C¶	D¶	
¶		

RTF Import New Features

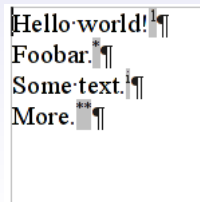
Footnotes

- All characters of the foot/endnote mark are in the field
- The field is properly superscripted
- Before:



A screenshot of a text editor showing four lines of text: "Hello world! 1", "Foobar.*", "Some text. 1", and "More.**". The numbers 1 and 2 are not superscripted, and the asterisks are not italicized. The text is enclosed in a white box with a thin border.

- After:



A screenshot of the same text editor showing the text after formatting. The numbers 1 and 2 are now superscripted, and the asterisks are italicized. The text is enclosed in a white box with a thin border.

RTF Import New Features

Line numbering

Before:

Lorem ipsum dolor
augue pellentesque
condimentum libero
leo. In dignissim
Maecenas malesuada
magna, at fringilla
ligula, condimentum
dignissim. Praesent
nisi ac dignissim.
pellentesque luct

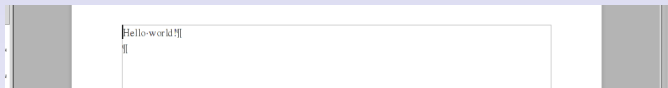
After:

1 Lorem ipsum dolor
augue pellentesque
condimentum libero
leo. In dignissim
5 Maecenas malesuada
magna, at fringilla
ligula, condimentum
dignissim. Praesent
nisi ac dignissim.
10 pellentesque luct

RTF Import New Features

Post-it fields

Before:



After:



RTF Import New Features

Form fields

Before:

```
This is a checkbox with help text: {}  
This one is selected: {}  
This is a textbox: Current text{}  
Listbox: {}
```

After:

```
This is a checkbox with help text: {}  
This one is selected: {}  
This is a textbox: Current text{}  
Listbox: Third entry{}
```

RTF Import New Features



Drawings

Before:

```
Rectangle:¶  
¶  
Ellipse:¶  
¶  
¶
```

```
Freeform line:¶  
¶  
¶
```

After:

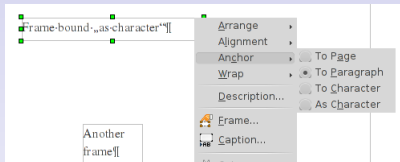
```
Rectangle:¶   
¶  
Ellipse:¶   
¶  
¶
```

```
Freeform line:¶   
¶  
¶
```

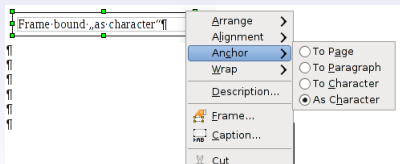
RTF Import New Features

Text frames

Before:



After:



DOCX Import Side-effects

Footnote restart, extra paragraphs

Before:

ctetur adipiscing elit. Maecenas condimentum consectetur est,
et. Suspendisse potenti. In hac habitasse platea dictumst. Praesent
cus volutpat. In vitae arcu sit amet velit commodo placerat in ut
ondimentum iaculis. Maecenas ac justo mauris, sit amet
seros. Suspendisse cursus, nulla in faucibus dictum, purus lorem
it ut tortor. Suspendisse potenti. Curabitur luctus malesuada nisi,
nullamcorper porta cursus. Quisque consequat semper nunc, eu
que mauris neque, facilisis sed molestie at, ultrices sed enim.
ut.

Vestibulum dapibus tincidunt cot
amet consectetur mauris. Quisque
turpis nec lectus rutrum rhoncus
ultrices consectetur varius. Cras
Nam vitae tortor a ligula interdum
vehicula nibh adipiscing.

1 Lorem ipsum dolor sit amet, consectetur adipiscing elit.

After:

tetur adipiscing elit. Maecenas condimentum consectetur est,
t. Suspendisse potenti. In hac habitasse platea dictumst. Praesent
is volutpat. In vitae arcu sit amet velit commodo placerat in ut
ndimentum iaculis. Maecenas ac justo mauris, sit amet
eros. Suspendisse cursus, nulla in faucibus dictum, purus lorem
ut tortor. Suspendisse potenti. Curabitur luctus malesuada nisi,
nullamcorper porta cursus. Quisque consequat semper nunc, eu
que mauris neque, facilisis sed molestie at, ultrices sed enim.
.

Vestibulum dapibus tincidunt cor
amet consectetur mauris. Quisque
turpis nec lectus rutrum rhoncus u
ultrices consectetur varius. Cras a
vitae tortor a ligula interdum aliq
vehicula nibh adipiscing.

1 Lorem ipsum dolor sit amet, consectetur adipiscing elit.

Acknowledgements, References

Thanks to – in no particular order:

- Cédric Bosdonnat and Björn Michaelsen: my mentors
- Luboš Lunak: writerfilter help
- Michael Stahl: initial tokenizer help
- Caolán McNamara: unit test
- Everyone else who helped on `#libreoffice-dev`

References:

- LibreOffice: <http://www.libreoffice.org/>
- SoC: <http://code.google.com/soc/>
- New RTF import filter:
<http://cgit.freedesktop.org/libreoffice/core/tree/writerfilter/source/rtftok>