Formatted Text Processing

In Docx DOM

By: Jarosław Kuchta

# Problem of structured text processing

When editing text in OpenXml Wordprocessing Data Object Model, we have a problem with the text structure. Text is stored in Paragraphs, which contain Run elements and other, non-textual elements. Run elements describe formatting of the textual elements which are contained in them.

Example: The following text:

This text is **bold,** and this is *italicized*.

is stored in OpenXml as:

<w:p w14:paraId="763D35DE" w14:textId="4BD8BFE4" w:rsidR="00B65346" w:rsidRDefault="006F2D84" w:rsidP="00307B9B">

<w:pPr>

<w:pStyle w:val="Standardowyakapit"/>

</w:pPr>

<w:r>

<w:t xml:space="preserve">This text is </w:t>

</w:r>

<w:r w:rsidRPr="006F2D84">

<w:rPr>

<w:b/>

<w:bCs/>

</w:rPr>

<w:t>bold,</w:t>

</w:r>

<w:r>

<w:t xml:space="preserve"> and this is </w:t>

</w:r>

<w:r w:rsidRPr="006F2D84">

<w:rPr>

<w:i/>

<w:iCs/>

</w:rPr>

<w:t>italicized</w:t>

</w:r>

<w:r>

<w:t>.</w:t>

</w:r>

</w:p>

The meanings of the XML elements are the following:

<w:p> - paragraph

<w:pPr> - paragraph properties

<w:pStyle> - paragraph style

<w:r> - run

<w:t> - text

<w:rPr> - run properties

<w:b> - bold

<w:bCs> - bold for complex script

<w:i> - italics

<w:iCs> - italics for complex script

The meanings of the XML attributes are the following:

w14:paraId – paragraph identifier

w14:textId – text identifier

w:rsidR – revision identifier for paragraph

w:rsidRDefault – default revision identifier for runs

w:rsidP – revision identifier for paragraph properties

w:rsidRPr – revision identifier for run properties

As we can see, there is plenty of excessive information: identifiers of paragraphs, runs, and revisions, properties of paragraphs and runs. This information would be lost if we simply get the text, edit it, and set back.

Example: Let’s try to change the text “bold" to “boldfaced" and “italicized" to “italic". To do it, first we get plain text from the structure. It is:

This text is bold, and this is italicized.

Next, we change the extracted text to:

This text is boldfaced, and this is italic.

Finally, we set the text back to the paragraph. In the naive solution, in the first step we remove all the member elements (except the paragraph properties).

<w:p w14:paraId="763D35DE" w14:textId="4BD8BFE4" w:rsidR="00B65346" w:rsidRDefault="006F2D84" w:rsidP="00307B9B">

<w:pPr>

<w:pStyle w:val="Standardowyakapit"/>

</w:pPr>

</w:p>

In the next step we create a new run with the changed text.

<w:p w14:paraId="763D35DE" w14:textId="4BD8BFE4" w:rsidR="00B65346" w:rsidRDefault="006F2D84" w:rsidP="00307B9B">

<w:pPr>

<w:pStyle w:val="Standardowyakapit"/>

</w:pPr>

<w:r>

<w:t xml:space="preserve">This text is boldfaced, and this is italic.</w:t>

</w:r>

</w:p>

We have lost the formatting, and the changed text looks simply as:

This text is boldfaced, and this is italic.

To preserve formatting, we must use another solution. Let’s take all run elements and store them with their text in the two-column list.

To identify the runs, we enter runId attributes to each run. These attributes are not defined in OpenXml and are shown below only for presentation purposes.

Example:

<w:p w14:paraId="763D35DE" w14:textId="4BD8BFE4" w:rsidR="00B65346" w:rsidRDefault="006F2D84" w:rsidP="00307B9B">

<w:pPr>

<w:pStyle w:val="Standardowyakapit"/>

</w:pPr>

<w:r runId="00000001">

<w:t xml:space="preserve">This text is </w:t>

</w:r>

<w:r w:rsidRPr="006F2D84" runId="00000002">

<w:rPr>

<w:b/>

<w:bCs/>

</w:rPr>

<w:t>bold,</w:t>

</w:r>

<w:r runId="00000003">

<w:t xml:space="preserve"> and this is </w:t>

</w:r>

<w:r w:rsidRPr="006F2D84" runId="00000004">

<w:rPr>

<w:i/>

<w:iCs/>

</w:rPr>

<w:t>italicized</w:t>

</w:r>

<w:r runId="00000005">

<w:t>.</w:t>

</w:r>

</w:p>

In the example, we have five runs with five strings, so we get the following table:

|  |  |
| --- | --- |
| <w:r runId="00000001"> | “This text is “ |
| <w:r runId="00000002"> | “bold," |
| <w:r runId="00000003"> | “ and this is “ |
| <w:r runId="00000004"> | “italicized" |
| <w:r runId="00000005"> | “." |

Let’s name the type of this table a FormattedText class as each string is associated with a Run element that holds formatting RunProperties.

Let’s define a GetText() method of the FormattedText which simply concatenates the strings to produce a plain text for the paragraph:

This text is bold, and this is italicized.

To change the text of the paragraph, we operate not on the result of the GetText() method, but on overall FormattedText structure.

Instead of calling GetText() and SetText() methods of the Paragraph class and Replace() method of the String class:

Paragraph.SetText(Paragraph.GetText().Replace(“bold", “boldfaced").Replace(“italicized", “italic"))

We create FormattedText instance on the Paragraph and invoke Replace() method of the FormattedText class:

var formattedText = new FormattedText(Paragraph);

formattedText.Replace(“bold", “boldfaced");

formattedText.Replace(“italicized", “italic");

The Replace(searchText, replacementText) method is implemented as follows:

var s = GetText();

var k = s.IndexOf(searchText);

if (k≥0)

ReplaceAt(k, searchText.Length, replacementText);

The ReplaceAt(position, length, replacementText) method iterates over FormattedText rows summarizing string lengths until we achieve a given position.

var sumLength = 0;

var selectedItem = -1;

for (int i=0; i≤ FormattedText.Count; i++)

{

var itemText = FormattedText[i].Text;

if (sumLength + itemText.Length > position)

{

selectedItem = i;

break;

}

sumLength += itemText.Length;

}

Then, in the selected string, it deletes a substring at index of position – sumLength with a given length and inserts the replacementText at this index. The modified text is set to the Run element.

if (selectedItem ≥ 0)

{

FormattedText[selectedItem].Text = FormattedText[selectedItem].Text.Remove(position – sumLength, length).Insert(position – sumLength, replacementText);

FormattedText[selectedItem].Run.SetText(FormattedText[selectedItem].Text);}

}

In our example we get FormattedText as follows:

|  |  |
| --- | --- |
| <w:r runId="00000001"> | “This text is “ |
| <w:r runId="00000002"> | “boldfaced," |
| <w:r runId="00000003"> | “ and this is “ |
| <w:r runId="00000004"> | “italic" |
| <w:r runId="00000005"> | “." |

Because the first column of the FormattedText refers to the Run elements contained in the Paragraph structure, we get:

<w:p w14:paraId="763D35DE" w14:textId="4BD8BFE4" w:rsidR="00B65346" w:rsidRDefault="006F2D84" w:rsidP="00307B9B">

<w:pPr>

<w:pStyle w:val="Standardowyakapit"/>

</w:pPr>

<w:r runId="00000001">

<w:t xml:space="preserve">This text is </w:t>

</w:r>

<w:r w:rsidRPr="006F2D84" runId="00000002">

<w:rPr>

<w:b/>

<w:bCs/>

</w:rPr>

<w:t>boldfaced,</w:t>

</w:r>

<w:r runId="00000003">

<w:t xml:space="preserve"> and this is </w:t>

</w:r>

<w:r w:rsidRPr="006F2D84" runId="00000004">

<w:rPr>

<w:i/>

<w:iCs/>

</w:rPr>

<w:t>italic</w:t>

</w:r>

<w:r runId="00000005">

<w:t>.</w:t>

</w:r>

</w:p>

and the formatting is preserved.

## More sophisticated problem

The above solution works well when we change the text which fits completely to a single Run. But what will happen when we change the text that is stored in more than one subsequent Run elements? We need to recode the change algorithm. Now it will delete only as much text in the selected FormattedText item as it fits in this item. The rest of the text will be deleted from the next item (or items). We will also introduce some auxiliary variables to debug the process of change.

if (selectedItem >= 0)

{

while (selectedItem < FormattedText.Count && (length > 0 || replacementText.Length > 0))

{

var itemText = FormattedText[selectedItem].Text;

var itemOldLength = itemText.Length;

var itemPosition = position - sumLength;

var itemRestLength = itemText.Length - itemPosition;

var delLength = length;

if (itemRestLength < length)

{

delLength = itemRestLength;

length -= delLength;

}

else

length = 0;

itemText = itemText.Remove(itemPosition, delLength);

itemText = itemText.Insert(itemPosition, replacementText);

replacementText = String.Empty;

FormattedText[selectedItem].Text = itemText;

FormattedText[selectedItem].Run.SetText(itemText);

if (length == 0)

break;

position += delLength;

sumLength += itemOldLength;

selectedItem++;

}

return true;

}

In the result, found text was properly deleted from subsequent Run elements and the replacement text was inserted to the first Run.

<w:p w14:paraId="763D35DE" w14:textId="4BD8BFE4" w:rsidR="00B65346" w:rsidRDefault="006F2D84" w:rsidP="00307B9B">

<w:pPr>

<w:pStyle w:val="Standardowyakapit"/>

</w:pPr>

<w:r>

<w:t xml:space="preserve">This text is </w:t>

</w:r>

<w:r w:rsidRPr="006F2D84">

<w:rPr>

<w:b/>

<w:bCs/>

</w:rPr>

<w:t xml:space="default">bold and this text is</w:t>

</w:r>

<w:r>

<w:t xml:space="preserve"> </w:t>

</w:r>

<w:r w:rsidRPr="006F2D84">

<w:rPr>

<w:i/>

<w:iCs/>

</w:rPr>

<w:t>italicized</w:t>

</w:r>

<w:r>

<w:t>.</w:t>

</w:r>

</w:p>

The formatted result looks like this:

This text is **bold and this text is** *italicized*.

The inserted text has got the formatting of the first found Run.