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="Normal"/>

</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:rsidRPr="006F2D84">

<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:rsidRPr="006F2D84">

<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.

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 will change the extracted text to:

This text is boldfaced, and this is italic.

Finally, we will 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="Normal"/>

</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="Normal"/>

</w:pPr>

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

<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 my:runId attributes to each run. These attributes (marked in green) are not defined in OpenXml and are shown below only for presentation purposes.

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

<w:pPr>

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

</w:pPr>

<w:r w:rsidRPr="006F2D84" my:runId="00000001">

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

</w:r>

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

<w:rPr>

<w:b/>

<w:bCs/>

</w:rPr>

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

</w:r>

<w:r w:rsidRPr="006F2D84" my:runId="00000003">

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

</w:r>

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

<w:rPr>

<w:i/>

<w:iCs/>

</w:rPr>

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

</w:r>

<w:r w:rsidRPr="006F2D84" my: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 my:runId="00000001"> | “This text is “ |
| <w:r my:runId="00000002"> | “bold," |
| <w:r my:runId="00000003"> | “ and this is “ |
| <w:r my:runId="00000004"> | “italicized" |
| <w:r my: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)

{

var itemPosition = position - sumLength;

FormattedText[selectedItem].Text = FormattedText[selectedItem].Text.Remove(itemPosition, length).Insert(itemPosition, replacementText);

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

}

In our example we get FormattedText as follows:

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

</w:pPr>

<w:r w:rsidRPr="006F2D84" my:runId="00000001">

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

</w:r>

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

<w:rPr>

<w:b/>

<w:bCs/>

</w:rPr>

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

</w:r>

<w:r w:rsidRPr="006F2D84" my:runId="00000003">

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

</w:r>

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

<w:rPr>

<w:i/>

<w:iCs/>

</w:rPr>

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

</w:r>

<w:r w:rsidRPr="006F2D84" my:runId="00000005">

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

</w:r>

</w:p>

and the formatting is preserved.

## Multiple runs 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, as below?

textProcessor.Replace(", and this is", " and this text is");

Notice that the comma character is contained in the “bold” Run, and the rest of the search text is contained in the next “no-bold” Run.

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.

We organized an internal while loop which starts at the selected FormattedText item and continues for the next items to delete the rest of text which does not fit to this item. The first part of the internal loop (in green) evaluates the length of the deletion. The last part of the loop (in red) is executed only when there is more text to delete, and it prepares the next iteration of the loop.

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;

if (delLength>0)

itemText = itemText.Remove(itemPosition, delLength);

if (replacementText.Length > 0)

itemText = itemText.Insert(itemPosition, replacementText);

FormattedText[selectedItem].Text = itemText;

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

if (length == 0)

break;

replacementText = String.Empty;

position += delLength;

sumLength += itemOldLength;

selectedItem++;

}

return true;

}

In our example, the found text was properly deleted from selected and subsequent Run elements and the replacement text was inserted to the selected Run.

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

<w:pPr>

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

</w:pPr>

<w:r w:rsidRPr="006F2D84" my:runId="00000001">

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

</w:r>

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

<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:rsidRPr="006F2D84" my:runId="00000003">

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

</w:r>

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

<w:rPr>

<w:i/>

<w:iCs/>

</w:rPr>

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

</w:r >

<w:r w:rsidRPr="006F2D84" my:runId="00000005">

<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.

## Problem of replacement text format

Now, how could we change the text and set it to some specific format?

First, we need to pass a specific format to the Replace() method. We cannot use a RunProperies element to represent the format for two reasons. First is that RunProperies element is not the only one description of the text formatting. The current formatting is applied using a *style hierarchy* in the following order:

1. DefaultRunProperties of the DocDefaults element,
2. ConditionalRunProperties of a table style,
3. RunProperties of a numbering style,
4. RunProperties of a paragraph style,
5. RunProperties of a character style,
6. RunProperties of the Run element

The second reason is that we could need some other formatting, like FontName, which is not directly contained in RunProperties.

So, we should declare a TextFormat class (or rather a record) with text attributes like Bold and Italic. The IsSame() method compares the text format with another text format. If some format attributes are not set, they are not compared.

public record TextFormat

{

public bool? Bold { get; set; }

public bool? Italic { get; set; }

public bool IsSame(TextFormat other)

{

if (Bold.HasValue && other.Bold.HasValue && Bold != other.Bold)

return false;

if (Italic.HasValue && other.Italic.HasValue && Italic != other.Italic)

return false;

return true;

}

}

We should also declare GetFormat() method of the Run element. It should have an implementation like:

public static TextFormat GetFormat(this DXW.Run run)

{

return new TextFormat()

{

Bold = run.IsBold(),

Italic = run.IsItalic()

};

}

We will need also a SetFormat() method with an implementation like this:

public static void SetFormat(this DXW.Run run, TextFormat format)

{

var runProperties = run.GetRunProperties();

if (format.Bold.HasValue)

{

runProperties.SetBold(format.Bold.Value, false);

runProperties.SetBold(format.Bold.Value, true);

}

if (format.Italic.HasValue)

{

runProperties.SetBold(format.Italic.Value, false);

runProperties.SetBold(format.Italic.Value, true);

}

}

We should declare a replacementFormat parameter of the TextFormat type in the Replace and ReplaceAt methods.

Replace(string searchText, string replacementText, TextFormat? replacementFormat = null);

ReplaceAt(int position, int length, string replacementText, TextFormat? replacementFormat = null);

For now, we must change the part of the code which inserts the replacementText.

If the replacementFormat is not null and is not same as the format of the selected item Run, the replacementText is not inserted to the selected Run, but the new Run is created and inserted.

First, we will save the text after deletion.

if (delLength > 0)

{

itemText = itemText.Remove(itemPosition, delLength);

FormattedText[selectedItem].Text = itemText;

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

}

There are three cases of new Run insertion:

* Before the selected Run – this is the case when item position of insertion is 0 (green code),
* After the selected Run –when item position is at the end of the item text (blue code),
* Inside the selected Run – then the selected Run must be split into two parts, and the new Run must be inserted between them (red code). We used the extension method SplitAt() of the Run element. This method divides the Run element in two parts at the given character position and returns the second part to the caller.

The violet code handles the case when there is no need to create a new Run.

var nextItem = selectedItem + 1;

if (replacementText.Length > 0)

{

if (replacementFormat != null && !replacementFormat.IsSame(FormattedText[selectedItem].Run.GetFormat()))

{

var selectedRun = FormattedText[selectedItem].Run;

if (itemPosition == 0)

{

var newRun = new DXW.Run();

newRun.AppendText(replacementText);

newRun.SetFormat(replacementFormat);

selectedRun.InsertBeforeSelf(newRun);

FormattedText.Insert(selectedItem, new RunText(newRun, replacementText));

}

else if (itemPosition == itemText.Length)

{

var newRun = new DXW.Run();

newRun.AppendText(replacementText);

newRun.SetFormat(replacementFormat);

selectedRun.InsertAfterSelf(newRun);

FormattedText.Insert(selectedItem + 1, new RunText(newRun, replacementText));

nextItem++;

}

else

{

var tailRun = selectedRun.SplitAt(itemPosition, TextOptions.PlainText);

if (tailRun != null)

{

selectedRun.InsertAfterSelf(tailRun);

selectedItem++;

FormattedText.Insert(selectedItem, new RunText(tailRun, tailRun.GetText(GetTextOptions)));

}

var newRun = new DXW.Run();

newRun.AppendText(replacementText);

newRun.SetFormat(replacementFormat);

FormattedText[selectedItem].Run.InsertAfterSelf(newRun);

FormattedText.Insert(selectedItem + 1, new RunText(newRun, replacementText));

nextItem++;

}

}

else

{

itemText = itemText.Insert(itemPosition, replacementText);

FormattedText[selectedItem].Text = itemText;

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

}

replacementText = String.Empty;

}

if (length == 0)

break;

position += delLength;

sumLength += itemOldLength;

selectedItem = nextItem;

}

In the example below, we can see a newly added Run (marked in red):

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

<w:pPr>

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

</w:pPr>

<w:r w:rsidRPr="006F2D84" my:runId="00000001">

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

</w:r>

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

<w:rPr>

<w:b/>

<w:bCs/>

</w:rPr>

<w:t xml:space="default">bold</w:t>

</w:r>

<w:r>

<w:rPr>

<w:b w:val="false"/>

<w:bCs w:val="false"/>

</w:rPr>

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

</w:r>

<w:r w:rsidRPr="006F2D84" my:runId="00000003">

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

</w:r>

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

<w:rPr>

<w:i/>

<w:iCs/>

</w:rPr>

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

</w:r>

<w:r w:rsidRPr="006F2D84" my:runId="00000005">

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

</w:r>

</w:p>

And the result is as follows:

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

## Searching text with a format

Now, we must change the example to show search text with a search format. We change the first “is” word to boldface and the next “is” to italic.

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

Let’s try to change the second “is”, which is italicized, and omit the first “is” (a bold one).

To search for a text with formatting, we need to add a searchFormat parameter to the Replace() method.

Replace(string searchText, TextFormat? searchFormat,   
string replacementText, TextFormat? replacementFormat)

Following this, we define a Search() method to get the position of text with format in the FormattedText starting at the given startPosition. The method returns the character position of the text, if found, or -1, if the text was not found. Text (and format) comparing must not be limited to the single Run, but it must be collected from the subsequent Runs (if the format is the same).

public int Search(int startPosition, string searchText, TextFormat? searchFormat)

{

var searchTextLength = searchText.Length;

var sumLength = 0;

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

{

var itemText = FormattedText[i].Text;

if (sumLength + itemText.Length > startPosition)

{

if (searchFormat == null || searchFormat.IsSame(FormattedText[i].Run.GetFormat()))

{

var textToCompare = itemText;

int j = i + 1;

while (textToCompare.Length > searchTextLength && j < FormattedText.Count &&

(searchFormat == null || searchFormat.IsSame(FormattedText[j].Run.GetFormat())))

{

textToCompare += FormattedText[j].Text;

j++;

}

itemText = FormattedText[i].Text;

if (textToCompare.Length >= searchTextLength)

{

var k = textToCompare.IndexOf(searchText);

if (k >= 0)

{

return sumLength + k;

}

}

}

}

sumLength += itemText.Length;

}

return -1;

}

The Search() method is used in the Replace() method as follows:

public bool Replace(string searchText, TextFormat? searchFormat,   
string replacementText, TextFormat? replacementFormat)

{

int k;

if (searchFormat != null)

k = Search(0, searchText, searchFormat);

else

{

var s = GetText();

k = s.IndexOf(searchText);

}

if (k >= 0)

{

ReplaceAt(k, searchText.Length, replacementText, replacementFormat);

return true;

}

return false;

}

We invoke the Replace() method requesting replace the “ is ” text with the format of italic by the text “ is ” with no italic. Spaces around “is” are needed to omit “is” in the word “this”.

textProcessor.Replace(" is ", new TextFormat { Italic = true },   
" is ", new TextFormat { Italic = false });

The result is as below. We can see two Runs with the same runId. The second one is created by the SplitAt() method. The newly created Run is marked in red.

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

<w:pPr>

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

</w:pPr>

<w:r w:rsidRPr="006F2D84" my:runId="00000001">

<w:rPr/>

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

</w:r>

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

<w:rPr>

<w:b/>

<w:bCs/>

</w:rPr>

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

</w:r>

<w:r w:rsidRPr="006F2D84" my:runId="00000003">

<w:rPr/>

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

</w:r>

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

<w:rPr>

<w:i/>

<w:iCs/>

</w:rPr>

<w:t xml:space="default">this</w:t>

</w:r>

<w:r>

<w:rPr>

<w:b w:val="false"/>

<w:bCs w:val="false"/>

</w:rPr>

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

</w:r>

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

<w:rPr>

<w:i/>

<w:iCs/>

</w:rPr>

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

</w:r>

<w:r w:rsidRPr="006F2D84" my:runId="00000005">

<w:rPr/>

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

</w:r>

</w:p>

The formatted result is:

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

## Search the whole words

We used a special trick to search the word “is” by surrounding the search text with spaces. We can introduce a boolean option of FindWholeWordsOnly. The Search() method must use this option in code.

public int Search(int startPosition, string searchText, TextFormat? searchFormat)

{

if (searchFormat == null)

{

var compareText = GetText();

if (FindWholeWordsOnly)

{

compareText += '\0' + compareText + '\0';

var k = compareText.IndexOf(searchText, startPosition);

while (k >= 0)

{

if (k > 0 && char.IsLetterOrDigit(compareText[k - 1]) || k + searchText.Length < compareText.Length && char.IsLetterOrDigit(compareText[k + searchText.Length]))

k = compareText.IndexOf(searchText, k + 1);

else

return k;

}

return -1;

}

else

return compareText.IndexOf(searchText, startPosition);

}

var searchTextLength = searchText.Length;

var sumLength = 0;

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

{

var itemText = FormattedText[i].Text;

if (sumLength + itemText.Length > startPosition)

{

if (searchFormat.IsSame(FormattedText[i].Run.GetFormat()))

{

var textToCompare = itemText;

if (FindWholeWordsOnly)

{

if (i > 0 && searchFormat.IsSame(FormattedText[i - 1].Run.GetFormat()))

textToCompare = FormattedText[i - 1].Text.LastOrDefault() + textToCompare;

else

textToCompare = '\0' + textToCompare;

}

int j = i + 1;

while (textToCompare.Length > searchTextLength && j < FormattedText.Count && searchFormat.IsSame(FormattedText[j].Run.GetFormat()))

{

textToCompare += FormattedText[j].Text;

j++;

}

if (FindWholeWordsOnly)

{

if (i < FormattedText.Count - 1 &&

searchFormat.IsSame(FormattedText[i + 1].Run.GetFormat()))

textToCompare = textToCompare + FormattedText[i + 1].Text.LastOrDefault();

else

textToCompare = textToCompare + '\0';

}

if (textToCompare.Length >= searchTextLength)

{

var k = textToCompare.IndexOf(searchText);

while (k >= 0)

{

if (FindWholeWordsOnly && k > 0)

{

if (char.IsLetterOrDigit(textToCompare[k - 1])   
|| char.IsLetterOrDigit(textToCompare[k + searchTextLength]))

k = textToCompare.IndexOf(searchText, k + 1);

else

return sumLength + k - 1;

}

else

return sumLength + k;

}

}

}

}

sumLength += itemText.Length;

}

return -1;

}

Replace() method can be simplified:

public bool Replace(string searchText, TextFormat? searchFormat,   
string replacementText, TextFormat? replacementFormat)

{

var k = Search(0, searchText, searchFormat);

if (k >= 0)

{

return ReplaceAt(k, searchText.Length, replacementText, replacementFormat);

}

return false;

}