Subject: Re: Another XML Question
Posted by Karl Schultz on Wed, 23 Mar 2005 17:16:00 GMT
View Forum Message <> Reply to Message

On Tue, 22 Mar 2005 20:52:12 -0700, David Fanning wrote:

> Michael Wallace writes:
>
>> So, when you say there's an "end of line string" or something in there,
>> what exactly are you talking about? I'm just curious what kind of
>> gremlin or banshee you're dealing with.
>
> Well, here is my XML file:
>
> <CONFIGDATA>
> <CAMPAIGN_ID>
> <TYPE>INT</TYPE>
> <VALUE>00</VALUE>
> </CAMPAIGN_ID>
> <SPAM_WAIT>
> <TYPE>INT</TYPE>
> <VALUE>60</VALUE>
> <UNITS>Seconds</UNITS>
> </SPAM_WAIT>
> </CONFIGDATA >
>
> Here is my code:
>
> doc = Obj_New('IDLffXMLDOMDocument')
> doc -> Load, Filename=filename, /Exclude_Ignorable_Whitespace
>
> tags = doc -> GetElementsByTagName('CONFIGDATA')
> node = tags -> Item(0)
> children = node -> GetChildNodes()
> FOR j=0,children->GetLength()-1 DO BEGIN
> child = children -> Item(j)
> Help, child
> ENDFOR
>
> And here is the result:
>
> CHILD OBJREF = <ObjHeapVar43440(IDLFFXMLDOMTEXT)>
> CHILD OBJREF = <ObjHeapVar43443(IDLFFXMLDOMELEMENT)>
> CHILD OBJREF = <ObjHeapVar43445(IDLFFXMLDOMTEXT)>
> CHILD OBJREF = <ObjHeapVar43447(IDLFFXMLDOMELEMENT)>
> CHILD OBJREF = <ObjHeapVar43449(IDLFFXMLDOMTEXT)>
>
> Only child 2 and 4 are the elements I'm looking for: CAMPAIGN_ID

> and SPAM_WAIT. The other three children are some kind of white
> space thingy. If I get the value of the TEXT objects, I see a single
> quote on one line and a single quote on the next line. No text.
>
> If I get the children of the CAMPAIGN_ID element, there are 9 of
> them, and only 3 I care about.
>
> Go figure!
>
> Cheers,
>
> David

In XML, whitespace is often considered significant, even in places where
you think it may not be.

For example,

<CAMPAIGN_ID>
 <TYPE>INT</TYPE>
 <VALUE>00</VALUE>
</CAMPAIGN_ID>

contains significant whitespace between the CAMPAIGN_ID start and end
tags. There are three newlines that correspond to the text nodes you
discovered above.

From an XML point of view, the above is QUITE different from:

<CAMPAIGN_ID><TYPE>INT</TYPE><VALUE>00</VALUE></CAMPAIGN_ID >

In this case, those three text nodes would not be in the DOM tree.

The XML folks wanted the whitespace to be detectable by the parser in case
there was an application need for that sort of information.

In order to teach the XML parser which whitespace is ignoreable and which
is not, you need to create a DTD or schema and specify the
EXCLUDE_IGNORABLE_WHITESPACE keyword. It is NOT sufficient to specify the
keyword without the DTD. (And the docs are *not* vague here :-).)

If you do make the DTD, you will not see those TEXT nodes containing
newlines or linefeeds. And the Windows vs Unix line terminator discussion
has nothing to do with this. You'd get the same result on either platform
and with either line terminator scheme. And that's by design.

If you don't want to make a DTD (it is worth doing, IMHO), you'd have to
beef up your parser to tolerate and skip over text nodes containing only

whitespace. Keep in mind too that an input file may contain:

<TYPE>INT</TYPE>

or

<TYPE>

INT

</TYPE>

and so you'd have to also deal with whitespace within an element where you might expect none.

If you make a DTD, your parser becomes MUCH simpler.

The IDL docs won't tell you how to make a DTD. For that, and for most other background XML expertise, you'll have to consult XML books, etc.

Karl
