

# **XML HACKS**

*100 Industrial-Strength Tips & Tools*



**O'REILLY**<sup>®</sup>

*Michael Fitzgerald*

HACK  
#38

## Pretty-Print XML Using a Generic Identity Stylesheet and Xalan

Sometimes your XML output from various programs is less than attractive. Spruce it up in a hurry with Xalan C++ and an identity transform.

In earlier hacks (“Edit XML Documents with Microsoft Word 2003” [Hack #14] and “Create an XML Document from a CSV File” [Hack #21]), you saw the unsightly XML output from Word 2003 and CSVToXML. The reason why this XML is unsightly is that it is output on only one or two lines. If you want this XML to be human-readable, here is a quick hack that pretty-prints the XML by properly indenting it.

In the working directory for this book you will find *identity.xml* (Example 3-13), a very simple identity stylesheet that effectively copies all nodes from source to the result as XML.

Example 3-13. *identity.xml*

```
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
<xsl:output method="xml" indent="yes" encoding="ISO-8859-1"/>

<xsl:template match="@*|node()">
  <xsl:copy>
    <xsl:apply-templates select="@*|node()"/>
  </xsl:copy>
</xsl:template>

</xsl:stylesheet>
```

The template matches all nodes (`node()`), including all attributes (`@*`), and copies them using the copy instruction, repeatedly applying templates until all nodes are processed. The `indent` attribute on the output element indents the output using a processor-dependent number of spaces. If you apply this stylesheet using Xalan C++ on the command line, you can use the `-i` switch to specify the number of spaces to indent the output. If you use Saxon, you can use the Saxon extension `saxon:indent-spaces` on output, and you can set the number of spaces used in indentation (use the namespace `xmlns:saxon="http://saxon.sf.net/"`). This makes it possible to turn *Time\_word.xml*, the WordprocessingML output of Word 2003, into something more readable.

Assuming that you have Xalan C++ downloaded and installed (from <http://xml.apache.org/xalan-c>), run this command at a shell prompt:

```
xalan -i 1 -o pretty.xml Time_word.xml identity.xml
```

## Pretty-Print XML Using a Generic Identity Stylesheet and Xalan

A sampling of *pretty.xml* is shown here, and it is much more readable than the original. It went from 2 long lines to 325 lines (Example 3-14).

Example 3-14. *pretty.xml*

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<?mso-application progid="Word.Document"?>

<w:wordDocument xmlns:w="http://schemas.microsoft.com/office/word/2003/wordml"
xmlns:v="urn:schemas-microsoft-com:vm1" xmlns:w10="urn:schemas-
microsoft-com:office:word" xmlns:s1="http://schemas.microsoft.com/schemaLibrary/
2003/core" xmlns:aml="http://schemas.microsoft.com/aml/2001/core" xmlns:wx="http:
//schemas.microsoft.com/office/word/2003/auxHint" xmlns:o="urn:schemas-microsoft-
com:office:office"
xmlns:dt="uuid:C2F41010-65B3-11d1-A29F-00AA00C14882"
w:macrosPresent="no" w:embeddedObjPresent="no" w:ocxPresent="no" xml:
space="preserve">
  <o:DocumentProperties>
    <o:Title>Time</o:Title>
    <o:Author>Mike Fitzgerald</o:Author>
    <o:LastAuthor>Mike Fitzgerald</o:LastAuthor>
    <o:Revision>2</o:Revision>
    <o:TotalTime>0</o:TotalTime>
    <o:Created>2004-02-11T23:07:00Z</o:Created>
    <o:LastSaved>2004-02-11T23:07:00Z</o:LastSaved>
    <o:Pages>1</o:Pages>
    <o:Words>9</o:Words>
    <o:Characters>52</o:Characters>
    <o:Lines>1</o:Lines>
    <o:Paragraphs>1</o:Paragraphs>
    <o:CharactersWithSpaces>60</o:CharactersWithSpaces>
    <o:Version>11.5604</o:Version>
  </o:DocumentProperties>
  <w:fonts>
    <w:defaultFonts w:ascii="Times New Roman" w:fareast="Times
New Roman" w:h-ansi="Times New Roman" w:cs="Times New Roman"/>
    <w:font w:name="Wingdings">
      <w:panose-1 w:val="05000000000000000000"/>
      <w:charset w:val="02"/>
      <w:family w:val="Auto"/>
      <w:pitch w:val="variable"/>
      <w:sig w:usb-0="00000000" w:usb-1="10000000"
w:usb-2="00000000" w:usb-3="00000000" w:csb-0="80000000"
w:csb-1="00000000"/>
    </w:font>
  </w:fonts>
```