

A Round Trip: from Word to XML and Back

Dorothy Hoskins & Terry Badger
Novatek XML Innovations Group

General Summary

Many content creation activities have reason to keep their content in a structured, open, and standard format. XML is now that de facto standard. The advantages are many, particularly when the content is delivered in different languages and different formats. Our discussion will be about a problem that our customer and many other organizations, encounter: some document sets are in XML and some are created and maintained in Microsoft Word. The solution needs to handle bi-directional format conversion between XML and Word. In our case, our customer wants to make XML the standard document format, but needs a method to convert Word documents to XML. In addition, our customer must provide some of their XML documents to end-users in a Word format so they can modify them for their own use.

We will demonstrate one solution which can be used as a model for other customers that have similar problems, using two commercial off-the-shelf (COTS) software applications, UpCast and XML Mind.

Detailed Summary

In brief, we will discuss our first step in the process using commercial off-the-shelf (COTS) software called UpCast. This will convert any Word document to an XML file based on the UpCast Document Type Definition (DTD). We will create style sheets to convert, using a general transformation (XSLT) of the XML file in UpCast format to the XML that matches the customer's standard XML DTD.

Normally this would be a one-time conversion, with the new XML files in the customer format becoming the file of record and the old Word document being archived and no longer updated.

Here are some of the obstacles encountered that will likely be faced by any organization contemplating a move from Word to XML:

1. Are standard Word styles used? A non-structured Word document can easily be structured by taking advantage of the heading styles.
2. Were features in tables used such as the column heading rows and vertical and horizontal cell spans? These can be used to present tables, even across page breaks, in a reasonable manner.
3. What symbols were used for footnote references? These references can be replicated and made into links to reference content in the XML.

4. How have images been referenced? Embedded image files in Word need special handling to become external references in XML.
5. Are there standard text strings that are ‘standard’ across all documents, but need to be changed to a new wording? The text change can be handled as part of the processing.
6. Are there ‘words’ that have special meaning, like a product code or a version date? Specific content can be captured in the up-translate process and be given special treatment to make them more useful in the XML output.
7. Is there boilerplate text that can be omitted but referenced and added later as needed? Some boilerplate is better handled as part of an output template rather than being included in each XML (company address and logo, disclaimers, copyright, etc.).
8. How are numbers and units of measure handled? These can be given US and non-US treatments as needed.
9. What special characters have been used from a symbol font and what do they mean? The processing can handle Greek/math symbols, special punctuation, etc. in a UTF-compliant manner.
10. What type and levels of lists are allowed? Nested lists can be captured as nested structures in the XML output.
11. How much use has been made of white space characters to make something ‘look right’, which may cause process problems? Stripping out tabs and line breaks, and/or replacing these with other special characters or spaces, can be part of the processing.
12. What use has been made of Word’s versioning control, field codes, bookmarks, TOC, indexes? Many of the Word-specific formatting can be captured and used to assist in creating rich XML output.
13. The “two of three” rule applies: usually you can get two of the three main customer goals: quality, price and speed, but there are tradeoffs to be determined by the customer’s priorities.

The second step of the process is to create a process that converts the customer XML to an intermediate file type, which then is passed to COTS software called XML Mind FO-Converter. XML Mind transforms the intermediate file format to produce a Word 2007 file format called .docx. (This format can be read directly by Word 2007 and converted to a .doc format for use by earlier version of Microsoft Word.)

This process can be used to create a separate user-editable Word version of the file when that is a format that is needed.

The intermediate .fo format can also be passed to a PDF rendering application such as Antenna House to create PDFs directly.

With these two processes the customer can move to the new paradigm of XML content creation without the loss of legacy Word documents.

Summary

Using COTS software and custom XSL templates, complex batch conversions from Microsoft Word to XML or XML to Microsoft Word can be performed in seconds using a standard PC.

References

- infinity-loop Home Document Conversion Software at <http://www.infinity-loop.de/>
- The Extensible Stylesheet Language Family (XSL) at <http://www.w3.org/Style/XSL/>
- XMLmind at <http://www.xmlmind.com/>
- SAXON The XSLT and XQuery Processor at <http://saxon.sourceforge.net/>
- Java at <http://www.java.com/en/download/>

Author Information

Dorothy Hoskins is XML Specialist for Novatek Communications of Rochester, NY and is "all XML, all the time". A frequent presenter on XML topics for the STC, Dorothy is also the author of "XML Publishing with InDesign CS2+" (O'Reilly Media, Inc. 2007 <http://oreilly.com/catalog/9780596513993/>). She primarily provides XSLT (transform) development, batch processing scripts and requirements gathering/quality assurance for XML projects involving HTML and PDF output.

Terry Badger is an XML Developer and has worked on a variety of XML projects over the course of 15 years. Terry has made a presentation at the STC and several presentations at the XML Extreme conferences: http://www.idealliance.org/papers/dx_xml03/papers/02-03-02/02-03-02.html and is currently teaching Introduction to XML at RIT. Terry's projects have included schema and DTD creation and process development using various commercial off-the-shelf (COTS) products and XSL stylesheets.