Paged Media with the DocBook xslTNG Stylesheets

Norman Tovey-Walsh

01 Jun 2023

A brief introduction to formatting paged media with the *DocBook xslTNG Stylesheets*.

(Follow along at home, https://nwalsh.com/scratch/markupuk2023/)

Table of Contents

What is DocBook?

DocBook is a very widely used XML vocabulary maintained by the DocBook Technical Committee (https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=docbook) of OASIS (https://www.oasis-open.org/). It is particularly well suited to books and papers about computer hardware and software (though it is by no means limited to these applications).

- Hierarchy: set, book, chapter, appendix, refentry, bibliography, glossary, index, …

- Blocks: section, para, table, figure, example, equation, programlisting, cmdsynopsis, classsynopsis, …

- Inlines: emphasis, filename, replaceable, foreignphrase, methodname, mousebutton, …

DocBook started as an SGML DTD more than 30 years ago. It is now an XML vocabulary defined normatively by a RELAX NG grammar and Schematron rules.

Styling DocBook

The *DocBook xslTNG Stylesheets* are the ~~third~~<ins>fifth</ins> generation of stylesheets that I've worked on for transforming DocBook into HTML and print. Before them there were:

- The *DocBook XSLT 2.0 Stylesheets*, circa 2010. (Paged media story: weak)

- The *DocBook XSLT 1.0 Stylesheets*, circa 2000 (still maintained). (Paged media story: XSL Formatting Objects)

- The *DocBook DSSSL Stylesheets*, circa 1996. (Paged media story: part of the processor)

- The *gmat* formatter (successor to *fmat*), circa 1994. (Paged media story: it generated troff markup from a recursive descent parse of the object tree. The less said, the better.)

  (It was a bailing wire, duct tape, and chewing gum hack that I co-wrote for O'Reilly. It used objects in some early beta of Perl 5 to build a parse tree of the SGML ESIS from DocBook (1.x? 2.x?) sources.)

What is DocBook xslTNG?

The *DocBook xslTNG Stylesheets* are XSLT 3.0 stylesheets for transforming DocBook to HTML. The goals are:

- Clean, semantic HTML

- That looks good "out of the box" (on the web or in print)

- With good accessibility hooks

- That is easy to style (in large part, perhaps, because of the efforts to generate clean HTML)

The standard distribution also includes extension functions to perform XInclude processing, syntax highlighting of verbatim environments, extract metadata from images, and even perform some validation.

Five Interesting Things

If you're interested in seeing more generally what the *DocBook xslTNG Stylesheets* can do online, see Five Interesting Things (https://xsltng.docbook.org/samples/fit/).

(P.S. There are at least seven interesting things now!)

Multi-stage transformation

Multi-stage transformation makes the processing simpler:

1. Cleanup the logical structure

2. Perform XInclude processing

3. (Convert DocBook 4.x to 5.x)

4. Perform DocBook Transclusion (https://docbook.github.io/documents/ transclusion.html)

5. Profile with effectivity attributes

6. Normalize markup

7. Process annotations

8. Apply XLink linkbases

9. (Validate with RELAX NG)

10. Transform to HTML (for online or paged media)

The last stage is always operating on consistent markup.

But wait, there's more

- The stylesheets can produce three different kinds of output: a single-page HTML document (the "natural" transformation of a DocBook document into an HTML document), a multi-page HTML document ("chunked" output), or HTML designed for use in paged media.

- As much as possible, we want to avoid caring about this in the DocBook-to-HTML transformation. The "core" xslTNG transformation is agnostic to the output type.

- Document features that need to be rendered differently (annotations, footnotes, etc.) are not in their final form in the initial HTML output.

- A small customization layer is provided for tailoring the output to paged media. This includes applying different CSS, for example.

- The HTML is transformed with a target in mind: for paged media, it generates HTML that can be styled appropriately for bottom-of-the-page footnotes, for example, and can add page numbers to cross references.

Example (DocBook XML)

A simple DocBook document:

```
 1  <article xmlns="http://docbook.org/ns/docbook"
            version="5.2">
    <info>
     <title>Hello, world</title>
 5   <author>
      <personname>Norm Tovey-Walsh</personname>
     </author>
     <copyright>
      <year>2023</year>
10    <holder>Norm Tovey-Walsh</holder>
     </copyright>
    </info>
    <para>This is a <emphasis>paragraph</emphasis>.<footnote>
     <para><quote>Pretty dull, huh?</quote></para></footnote>
15  </para>
   </article>
```

Example (online HTML)

The online HTML output

```
 1  <body class="home">
      <nav class="top"/>
      <main>
        <article class="article component">
 5        <header>
            <h1>Hello, world</h1>
            <div class="author">
              <h3>
                <span class="first-last personname">Norm Tovey-Walsh</span
10            </h3>
            </div>
            <div class="copyright">Copyright © <span class="copyright-years">
              <span class="year">2023</span>
            </span>
15            <span class="holder">Norm Tovey-Walsh</span>
            </div>
          </header>
          <p>This is a <em>paragraph</em>.<sup id="R_para1_footnote1-fref
            <a href="#R_para1_footnote1-fnote">1</a>
20          </sup>
          </p>
        </article>
        <footer>
          <div class="footnotes">
25          <hr/>
            <div id="R_para1_footnote1-fnote" class="footnote">
```

```
              <div class="footnote-number">
                <sup class="footnote-number">
                  <a href="#R_para1_footnote1-fref">1</a>
30              </sup>
              </div>
              <div class="footnote-body">
                <p>
                  <q>Pretty dull, huh?</q>
35              </p>
              </div>
            </div>
          </div>
        </footer>
40    </main>
      <nav class="bottom"/>
    </body>
```

Example (paged-media HTML)

The paged media HTML output

```
1  <body class="article component">
    <header>
      <h1>Hello, world</h1>
      <div class="author">
5       <h3>
          <span class="first-last personname">Norm Tovey-Walsh</span>
        </h3>
      </div>
      <div class="copyright">Copyright © <span class="copyright-years">
10        <span class="year">2023</span>
        </span>
        <span class="holder">Norm Tovey-Walsh</span>
      </div>
    </header>
15  <main>
      <p>This is a <em>paragraph</em>.<span id="R_para1_footnote1-fnote
        <q>Pretty dull, huh?</q>
      </span>
    </p>
20  </main>
  </body>
```

Example (PDF)

Sample PDF output. (This presentation (docbook.pdf) is also available as a PDF.)

Hello, worl

Norm Tovey-Walsh

This is a *paragraph*.[1]

A quick tour through some customizations

This presentation is a DocBook book with custom stylesheets and custom CSS.

Let's take a quick tour…

(Special thanks to Tony who's identified a number of bugs in the paged media support recently. *Most* of them are now fixed.)

Colophon

This book was authored in DocBook 5.2 with Emacs. It was formatted for presentation with the *DocBook xslTNG Stylesheets* version 2.1.4 using SAXON HE 11.5. It was formatted as PDF by Antenna House.