MENTEA

13 Kelly's Bay Beach
Skerries, Co. Dublin, Ireland

Phone/Fax: +353.1.810.6797
Skype: MenteaXML

info@mentea.net
www.mentea.net

Transformation As Well As Styling, Tony Graham, Mentea

Publishing and the Open Web Platform Workshop

Mentea is a Dublin-based consultancy specialising in XML, XSLT, and XSL-FO. Tony Graham has over 20 years experience working with markup in four countries across three continents. He is a member of the Japanese Layout Task Force, Chair of the Print and Page Layout Community Group, and author of Unicode: A Primer.

In many areas of publishing, the full text of the source is not the same as the text of the formatted result (or, in many cases, of any one of multiple formatted results), particularly when the source is the eventual archive format for the text. I will provide some examples at the end from scientific journal publishing, but transforming text for display is common to many publishing processes.

Transformations of markup doesn't usually get a mention when talking about the Open Web Platform [4][5]. To the extent that it's thought about at all, there's generally an assumption that it's-so-obvious-it-doesn't-need-to-be-stated that it's one of the things for which you'd use ECMAScript/JavaScript. However, just as doing something declaratively in CSS is typically seen as preferable to doing it with JavaScript programming, using a declarative language for transforming markup can be shorter, easier to write, and easier to understand than writing a functional program to do the same thing.

Bert Bos's position paper and presentation [1] for the eBooks and I18n workshop [2] in Tokyo in June allude to the possibility of using XSLT, or something like XSLT, as well as CSS. I couldn't go, and I haven't seen any minutes, so I don't know what reception it got there.

XSLT 1.0 natively runs, more or less, in major browsers but XSLT in the browser hasn't received much care and attention in recent years, and no browser vendor's implementation covers XSLT 2.0, which is more powerful and allows much more concise stylesheets than is typically possible with XSLT 1.0. Although XSLT 2.0 doesn't come in any browser, using it is possible today since Saxonica has a 'Saxon-CE' open source XSLT 2.0 processor boiled down to JavaScript that runs entirely in the browser [3].

Menteith Consulting Ltd trading as Mentea.
Registered Office: 13 Kelly's Bay Beach, Skerries, Co. Dublin.
Registered in Ireland, No. 428599, 427521. Directors: Tony Graham, Mary Tan (USA).

Wholesale adoption of XSLT, with its XML syntax, into the Open Web Platform would be unlikely. To develop an alternative compact, non-XML syntax for XSLT would be a bonus even for existing XSLT users, but beyond syntax, people will also point to differing data models or preferences for CSS selectors as reasons to not just adopt XSLT and its underlying XPath. However, if the necessity for transformation as well as styling is acknowledged, to ignore lessons learnt in developing XSLT and attempt to develop a new transformation language from scratch would just be making more work.

Transformation examples

These examples are from formatting XML of scientific journal articles marked up to the NISO Journal Archiving and Interchange, (JATS) version 1.0 schema [6] into two-column, paginated PDF using XSLT, XSL-FO 1.1, and some vendor extensions.

Author initials

The source XML and the formatted output both contain the authors' full names, but the formatted output also includes a sample citation listing the article title, the family names and initials of the first five authors, and other information. Beyond selectively reusing and reordering content, inserting punctuation, and simply taking the first character of an author's given name, any hyphenated names such as, e.g., "John-Jacques" need to show the initial from each name, e.g., "J-J".

Merge bibliographic references

Cross-references to citations in the bibliography are shown as numerals in square brackets, e.g., "[1]", but consecutive cross-references, rather than being shown as "...the tree frog is green [2],[3],[5]..." needed to be shown as "the free frog is green [2,3,5]...", and while consecutive cross-references to a consecutive range of citations are included individually in the source XML, more than two references in a row need to be presented as a range, e.g., "the free frog is green [2-5]...".

Rotated and multi-page tables

Depending on the content, tables need to be automatically formatted as column-wide, page-wide, or rotated to be page-high and either column-deep or page-deep and made to float to the top of their column or page. When a table is too tall to fit on one page, since floats in XSL 1.1 don't break across pages, the over-long tables have to be split into multiple tables, each in their own float.

Column-wide floats aren't part of XSL 1.1, so they require a vendor extension, and determining each table's optimum width and whether or not it is too tall for a page requires an initial formatting run that formats each table at each possible width and using the 'area tree' from that as a secondary input to the final stylesheet so it can make the right choices.

[1] http://www.w3.org/Talks/2013/0604-CSS-Tokyo/
[2] https://www.w3.org/2013/06/ebooks/program.php
[3] http://www.saxonica.com/ce/index.xml
[4] http://www.w3.org/2010/Talks/0117-next-web-plh/nextweb.html
[5] https://en.wikipedia.org/wiki/Open_Web_Platform
[6] http://jats.nlm.nih.gov/archiving/tag-library/1.0/index.html