**WWN Development Document**

MS Word’s HTML : description, features, and problems

Word’s Navigation pane shows the table-of-contents (View : Show : Navigation pane).

This document was created by the WWN author for his own use in developing WWN. It is included in the WWN repo, as other developers may find it useful.

# Word’s HTML: description and specs

* Cruft inside Microsoft Word HTML files - Zoompf Web Performance
  + <https://zoompf.com/blog/2010/01/cruft-inside-microsoft-word-html-files/>
  + *Word includes an option that allows you to* ***save a filtered HTML file****. A filter HTML file will not contain any of this useless Microsoft Office specific data.*
* What makes Microsoft-Word-generated HTML documents so large in code? - Stack Overflow
  + <https://stackoverflow.com/questions/30437400/what-makes-microsoft-word-generated-html-documents-so-large-in-code>
* HTML and MS Word HTML Output | Metanorma
  + <https://www.metanorma.com/builder/ref/html/>
  + Describes Word's HTML structure
* About Microsoft Word HTML | Metanorma
  + <https://www.metanorma.com/builder/ref/ms-word-html/>
  + Describes limitations of Word's HTML
* *a Microsoft document called "Microsoft Office HTML and XML Reference". It's a Windows .exe that unpacks to a .chm Help file. You can get it here*
  + Word html format: insert a custom TOC via field code - Stack Overflow
    - <https://stackoverflow.com/questions/59689248/word-html-format-insert-a-custom-toc-via-field-code>
  + On my PC: D:\Documents\Professional-projects\My-web-site-development\Word-to-HTML\references\MS-Word-HTML-specs

# Word's document-properties and HTML

* Summary
  + I looked into which Word properties are included in the generated HTML
    - My hypothesis was that the Word properties could be set via VBA, as part of a script to generate HTML for Word document.
    - And, those Word properties would be put in HTML, e.g., in meta tags
  + Experiment
    - I set all of the properties on Word's Info page
    - Title and Hyperlink Base might have been the only one that made it into the HTML
  + Summary
    - I didn't keep good records of what properties got into the HTML, but this approach wasn't fruitful, except for Title and Hyperlink Base
* **Document property**: Hyperlink Base
  + Specify a file path
  + For relative hyperlinks, in generated HTML
    - By default, it's relative to location of document
    - If Hyperlink Base specified, it's relative to that path
  + How to create absolute hyperlinks and relative hyperlinks in Word documents
    - <https://support.microsoft.com/en-us/topic/how-to-create-absolute-hyperlinks-and-relative-hyperlinks-in-word-documents-d6339b5a-98ba-a483-bac5-95bd33ff5a80>
  + (1) How can I make a Word document with hyperlinks portable using the Hyperlink Base? - Quora
    - <https://www.quora.com/How-can-I-make-a-Word-document-with-hyperlinks-portable-using-the-Hyperlink-Base>
* Document properties (meta-data)
  + html - How many keywords are ideal for the META keywords tag? - Stack Overflow
    - <https://stackoverflow.com/questions/3812143/how-many-keywords-are-ideal-for-the-meta-keywords-tag>
    - *As a general rule, you should aim for the following character limits within each of your meta tags:*
      * *Page title – 70 characters*
      * *Meta description – 160 characters*
      * *Meta keywords – No more than 10 keyword phrases*
  + View or change the properties for an Office file
    - <https://support.microsoft.com/en-us/topic/view-or-change-the-properties-for-an-office-file-21d604c2-481e-4379-8e54-1dd4622c6b75>
  + How to Add Tags to Word Documents
    - <https://www.lifewire.com/use-tags-to-organize-word-documents-3540109>
* Experiments
  + Files are in the directory: Word-to-HTML-experiments
    - helloworld.docx
      * All Word properties set
    - helloworld--filtered.htm
      * MS Word, saved as Web, filtered
    - helloworld.htm
      * MS Word, saved as html

# Investigating for Word features that do not render correctly to HTML

## Need to look into

* Sources, may be useful
  + Advice – Word to HTML
    - Word to HTML – Convert your text to clean code
      * <https://wordtohtml.net/content/>
      * NOTE: headings not rendered correctly for D:\Documents\Professional-projects\My-web-site-development\Word-to-HTML\Word-to-HTML-experiments\investigation--bullet-symbols\headings-and-bullets-01.docx
  + Accessibility at Penn State | Cautions on Converting Word to HTML
    - <https://accessibility.psu.edu/microsoftoffice/microsoftword/wordhtml/>
  + Convert Word to HTML
    - <https://ww2.nscc.edu/lyle_l/Accessibility/convertWordtoHTML.html>
  + <https://www.ferris.edu/it/howto/pdfs-docs/content_in_word.pdf>
    - Describes common formatting problems
  + <https://word.uservoice.com/>

## Problems found in regression-test Word docs

* Multiple columns in Word doc
  + The columns don’t seem to get implemented in Word HTML
  + In the Word HTML, there is just one column
  + An example is in the regresssion-testing files:
    - See: WWN--testing.docx
      * See test file: computer-concepts--original.html
* For the ordered list, the text after the list-symbol is slightly misaligned. It is indented too much, by about 1 character-width. This is a tolerable error, for now.
  + An example is in the regresssion-testing files:
    - See: WWN--testing.docx
      * See test file: physics-tutorial--original.docx

## Known problems in Word's HTML

* Limitations when you save a Word document as a Web page
  + <https://support.microsoft.com/en-us/topic/limitations-when-you-save-a-word-document-as-a-web-page-f361de08-ca4c-bc53-11ef-138c0e405c44>
  + Note: not useful, except for the list below
  + The following tables list the elements that Word changes or removes when it saves a file as a Web page.
    - Character Formatting
    - Paragraph Formatting
    - Page Layout
    - Editing
    - Graphic Formatting
    - Asian Text Formatting
    - Table Formatting
    - Web Page Formatting
* Accessibility at Penn State | Cautions on Converting Word to HTML
  + <https://accessibility.psu.edu/microsoftoffice/microsoftword/wordhtml/#sample>
  + *Accessibility and Usability Issues of Word Generated HTML*
    - Note: limitations described

# Problematic Word features to test

* Page break
  + <https://stackoverflow.com/questions/8218377/rendering-page-breaks-in-word-with-html>
* Tabs
  + <https://stackoverflow.com/questions/31031996/how-to-fix-word-html-tabs>
* Smart quotes (turn off)