April 18, 2017

Abstracting Structure of HTML Document to Optimize Presentation

Andrew Sacamano

Abstracting Structure of HTML Document to Optimize Presentation

Abstract:

A computer program can generate a simplified Hypertext Markup Language (HTML) file from an HTML source file and a style guide. The HTML source file may have been generated from a word processing file, and may also include Cascading Style Sheets (CSS) elements. The simplified HTML file can include presentation elements in a consistent format.

Some word processors present a what-you-see-is-what-you-get (WYSIWYG) interface that presents text and other elements of a document in the same manner that the text will appear on a printed page. Users can apply formatting to elements of the document to create a desired appearance.

In some applications, such as machine processing, it can be useful to present a document with indications of structural information. The structural information can then be used to present the document in different formats, such as for United States letter size, A4, e-readers, mobile devices, or desktop computers with varying display sizes. The structural information can indicate, for example, what part of a document (such as a section header) a particular passage of text comprises. While some word processors can convert documents into structured formats, such as Hypertext Markup Language (HTML) or eXtensible Markup Language (XML), the files resulting from these conversions can be difficult to interpret.

To abstract a structure to generate a structured document that optimizes presentation, a computer program can identify elements of an HTML file, determine the presentation styles of the identified elements based on a style guide, and create a structured HTML document with the identified elements in a format indicated by the style guide. The resulting HTML document will clearly indicate portions of the document in a manner that is easy for a user to understand.

FIG. 1 is a flowchart showing a method 100 for generating a structured document. The method 100 may begin with a document, which may have been generated by a word processor application, where the document includes source text such as HTML and/or Cascading Style

Sheets (CSS). The method 100 can also begin with a user-supplied reference style sheet, style guide, or canonical style sheet (referred to as a style guide hereinafter), which can be written in a structured language such as CSS. The style guide can include a restricted set of CSS that applies only styles that are specified in the style guide.

FIG. 2 is an example of an original document presented by a word processor in WYSIWYG format. In this example, the document includes a body 202, a title 204, a subtitle 206, a run of text 208 in italics, and a run of text 210 in bold.

FIG. 3 is an example of a style guide. The style guide may have been written by a user. The style guide indicates portions of text and styles that correspond to each portion of text. In this example, a body 302 will be considered the portion of text that is left-aligned, has a font size of twelve points, and has no left margin. A title 304 will be considered the portion of text that is center-aligned, has a font size of twenty-four points, and has no text indentation. A subtitle 306 will be considered the portion of text that is center-aligned, has a font size of sixteen points, has a font style of italic, and has no text indentation. Metadata 312 will be considered portions of text that are left aligned and have no indentations or margins. Paragraphs 314 will be considered text portions with indentations of three em's (em's are distance measurements determined based on font sizes). List items 316 will be considered text portions with left margins of three em's. Spans of italic text 308 will be considered text portions with a font style of italic. Spans of bold text 310 will be considered text portions with a font weight of bold.

FIGs. 4A, 4B, and 4C show an HTML document generated by a word processor from the original document shown in FIG. 2. As shown in FIG. 4A, 4B, and 4C, the HTML document is difficult to interpret, and has many extraneous elements. As shown in FIG. 4C, the HTML document includes the body 402 corresponding to the body 202 in the original document shown

in FIG. 2, the title 404 corresponding to the title 204 in the original document shown in FIG. 2, the subtitle 406 corresponding to the subtitle 206 in the original document shown in FIG. 2, italic text 408 (identified as italic by the indicator "T2," which is defined as italic toward the bottom of FIG. 4B) corresponding to the italic text 208 in the original document shown in FIG. 2, and bold text 410 (identified as bold by the indicator "T1," which is defined as bold toward the bottom of FIG. 4B) corresponding to the bold text 210 in the original document shown in FIG. 2.

The method 100 can include parsing the source text (shown in FIGs. 4A, 4B, and 4C) and a style guide (shown in FIG. 3) (102). The parsing can break the source text into HTML elements, and extract elements of the source text, such as divs, spans, and a body of text. The parsing can also extract the elements of the style guide that will be presented in the output document.

The method 100 can also include identifying blocks and runs (104). The blocks can occupy locations on a page, and the spans can be contiguous streams of characters within a single block.

The method 100 can include determining presentation styles of each of the blocks and runs (106). The presentation styles can be determined from the HTML and/or CSS of the HTML and/or CSS document generated by the word processor (an example of which is shown in FIGs. 4A, 4B, and 4C). The presentation styles can include alignments, indentation, font size, and font style (such as bold and/or italic) of text.

The method 100 can include finding closest matches of the determined presentation styles for each of the blocks and runs (108). Finding closest matches includes determining which of the elements in the style guide, an example of which is shown in FIG. 3, each of the blocks and runs

is most similar to. The similarity can be based on a weighted average of features such as font size, margin, alignment, and whether the text is in italic or bold style.

In some examples, to account for different font sizes used in documents, such as to accommodate readers who need larger fonts to read without difficulty, the method can include normalizing the font sizes before finding the closest matches. The font sizes can be normalized by a) comparing the font sizes present in the original document and using a heuristic (such as determining the most common font size) to identify a "baseline" font size in the original document, and b) performing a function, such as linear scaling, to scale the font sizes in the original document so that the baseline font size is mapped to the body font size in the style guide.

The method 100 can include applying the closest matches to the styles and runs (110). Applying the closest matches can include applying the formatting of the matching element in the style guide to each of the styles and runs. For example, while the "P10" indicator preceding the title 406 in the HTML document shown in FIG. 4C is defined in FIG. 4B as having a font size of twenty-eight points, the style guide of FIG. 3 prescribes twenty-four point font for the title 304, which will result in the title having a font size of twenty-four points. Other elements may have their styles changed to conform to the styles in the style guide.

The method 100 can include removing logically unnecessary structural elements (112). For example, "<i>one <\i> <i>two<\i>," which, when rendered, results in "*one two*," can be converted to "<i>one two<\i>," which results in equivalent output when rendered. The logically unnecessary structural elements can be removed by applying heuristic rules, such as, in this example, removing an end of a particular style (italic) that is immediately followed by a beginning of the same style (italic).

The method 100 can include removing stylistically unnecessary elements (114). Removing the stylistically unnecessary elements can include applying heuristics, such as removing leading and trailing whitespace from paragraphs.

The method 100 can include outputting a file (116). The outputted file can be the HTML file resulting from performing the above functions. An example of the outputted file is shown in FIG. 5. As shown in FIG. 5, the outputted file includes a body 502 corresponding to the body 202 in the original document shown in FIG. 2, a title 504 corresponding to the title 204 in the original document shown in FIG. 2, a subtitle 506 corresponding to the subtitle 206 in the original document shown in FIG. 2, italic text 508 corresponding to the bold text 208 in the original document shown in FIG. 2, and bold text 510 corresponding to the bold text 210 in the original document shown in FIG. 2. The document shown in FIG. 5 is easy for a user to read, and easy for a computer program to parse and render in various formats, such as United States letter size, A4, e-readers, mobile devices, or desktop computers with varying display sizes.

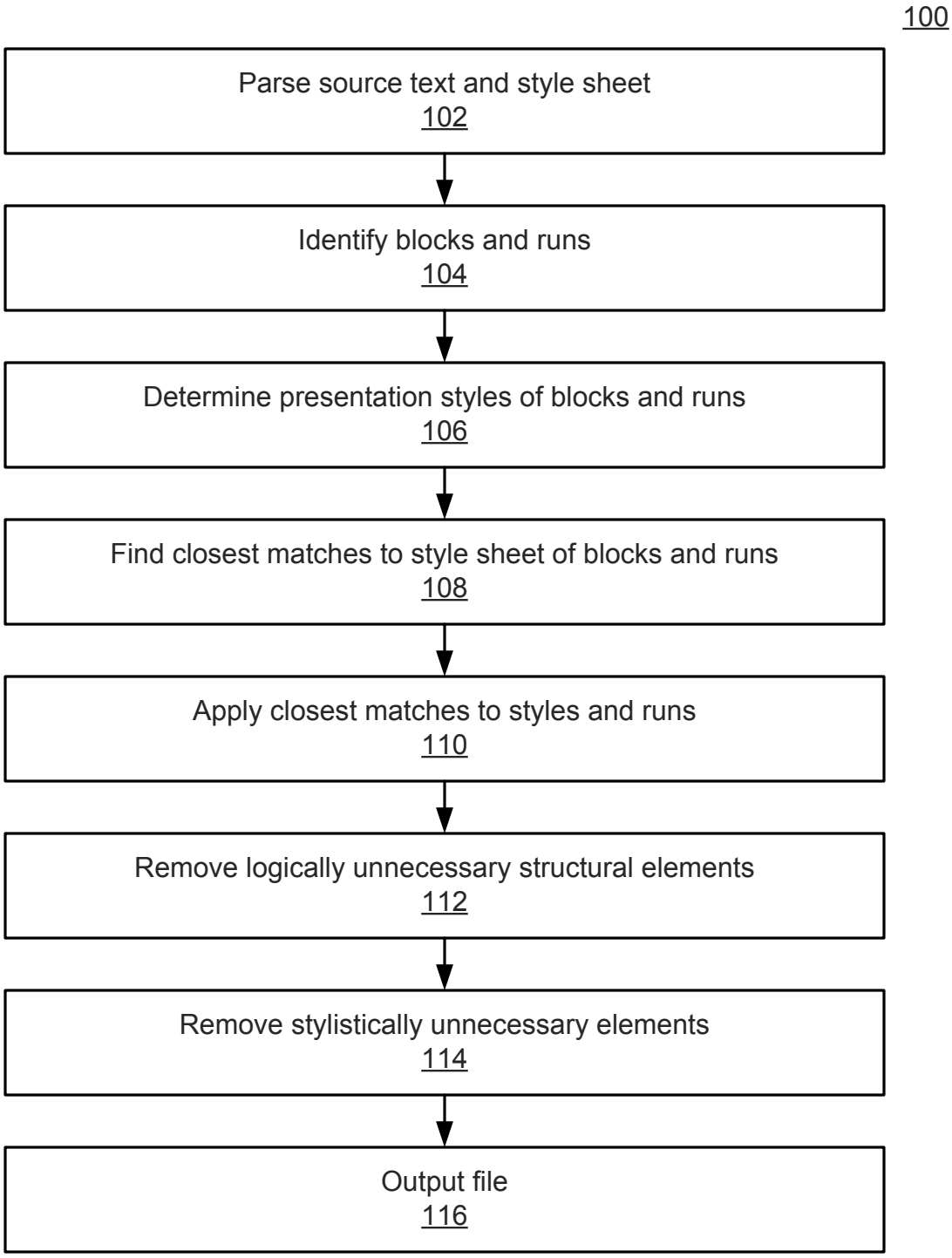
100
Parse source text and style sheet
102
Identify blocks and runs
104
Determine presentation styles of blocks and runs
106
Find closest matches to style sheet of blocks and runs
108
Apply closest matches to styles and runs
110
Remove logically unnecessary structural elements
112
Remove stylistically unnecessary elements
114
Output file
116

FIG. 1

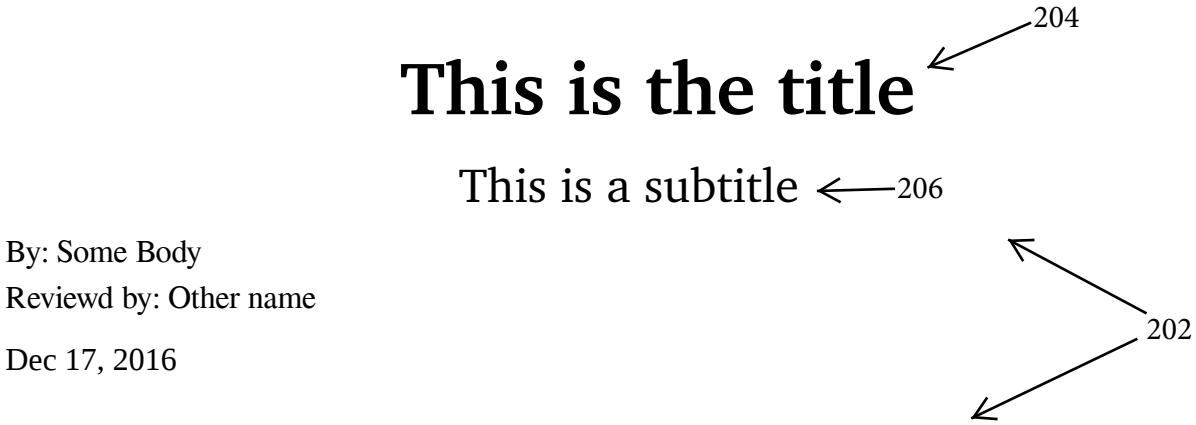
204
This is the title
This is a subtitle
206
By: Some Body
Reviewd by: Other name
Dec 17, 2016
202
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed scelerisque purus non elit ultricies, eu volutpat magna ullamcorper. Fusce viverra semper magna eget ultrices. Sed dolor tellus, fermentum sed aliquet ut, mattis in augue.
210
Mauris est risus, mattis at justo pretium, vestibulum posuere nibh. Nam mattis pellentesque neque, at imperdiet orci tristique vitae. Sed eget gravida leo, ut convallis leo. Suspendisse eget dui faucibus..
208
Nunc ac accumsan arcu. Aliquam erat mi, viverra vitae eros at, ultricies accumsan nisi. Duis volutpat eleifend augue in ultricies. Praesent sit amet neque mattis, viverra dolor quis, viverra purus. Mauris urna purus, rhoncus non metus id, lobortis congue felis. Morbi iaculis, quam sed porttitor aliquam, tellus eros porta orci, eget bibendum ante ipsum nec elit. Integer tempor ultrices leo, non varius lectus volutpat quis.
1. This is a list item
2. This is another list item
3. This is also a list item, but not really.
Sed gravida purus id justo placerat eleifend sed a justo. Nunc ullamcorper rhoncus tellus, sed aliquet sem euismod eget. Ut feugiat suscipit nibh, id vulputate purus tincidunt nec. Mauris in diam eget augue pharetra ornare sed condimentum lacus. Mauris semper vulputate elit eu consequat. Quisque scelerisque tempor nulla,

FIG. 2

styleguide.css

```
body {
  text-align: left;
  font-size: 12pt;
  margin-left: 0px;
}

p.title {
  text-align: center;
  font-size: 24pt;
  text-indent: 0em;
}

p.subtitle {
  text-align: center;
  font-size: 16pt;
  font-style: italic;
  text-indent: 0em;
}

p.metadata {
  text-align: left;
  text-indent: 0em;
  margin-top: 0px;
  padding-top: 0px;
  margin-bottom: 0px;
  padding-bottom: 0px;
}

p {
  text-indent: 3em;
}

li {
  margin-left: 3em;
}

span.italic {
  font-style: italic;
}

span.bold {
  font-weight: bold;
}
```

302 304 306 312 314 316 308 310

FIG. 3

DocFromWordProcessorAsHtml.html

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.1 plus MathML 2.0//EN"
"http://www.w3.org/Math/DTD/mathml2/xhtml-math11-f.dtd"><html
xmlns="http://www.w3.org/1999/xhtml"><!--This file was converted to xhtml by
LibreOffice - see
http://cgit.freedesktop.org/libreoffice/core/tree/filter/source/xslt for the
code.--><head profile="http://dublincore.org/documents/dcmi-terms/"><meta
http-equiv="Content-Type" content="application/xhtml+xml; charset=utf-8"/><title
xml:lang="en-US">- no title specified</title><meta name="DCTERMS.title" content=""
xml:lang="en-US"/><meta name="DCTERMS.language" content="en-US"
scheme="DCTERMS.RFC4646"/><meta name="DCTERMS.source"
content="http://xml.openoffice.org/odf2xhtml"/><meta name="DCTERMS.creator"
content="Andrew Sacamano"/><meta name="DCTERMS.issued"
content="2017-03-29T13:35:36.576181709" scheme="DCTERMS.W3CDTF"/><meta
name="DCTERMS.contributor" content="Andrew Sacamano"/><meta name="DCTERMS.modified"
content="2017-03-29T13:48:09.129729788" scheme="DCTERMS.W3CDTF"/><meta
name="DCTERMS.provenance" content="" xml:lang="en-US"/><meta name="DCTERMS.subject"
content="," xml:lang="en-US"/><link rel="schema.DC"
href="http://purl.org/dc/elements/1.1/" hreflang="en"/><link rel="schema.DCTERMS"
href="http://purl.org/dc/terms/" hreflang="en"/><link rel="schema.DCTYPE"
href="http://purl.org/dc/dcmitype/" hreflang="en"/><link rel="schema.DCAM"
href="http://purl.org/dc/dcam/" hreflang="en"/><style type="text/css">
        @page {  }
        table { border-collapse:collapse; border-spacing:0; empty-cells:show }
        td, th { vertical-align:top; font-size:12pt;}
        h1, h2, h3, h4, h5, h6 { clear:both }
        ol, ul { margin:0; padding:0;}
        li { list-style: none; margin:0; padding:0;}
        <!-- "li span.odfLiEnd" - IE 7 issue-->
        li span. { clear: both; line-height:0; width:0; height:0; margin:0;
padding:0; }
        span.footnodeNumber { padding-right:1em; }
        span.annotation_style_by_filter { font-size:95%; font-family:Arial;
background-color:#fff000;  margin:0; border:0; padding:0;  }
        * { margin:0;}
```

FIG. 4A

DocFromWordProcessorAsHtml.html

```
.P1 { font-size:12pt; line-height:120%; margin-bottom:0.0972in; margin-top:0in;
font-family:Liberation Serif; writing-mode:page; }
        .P10 { font-size:28pt; font-weight:bold; margin-bottom:0.0835in;
margin-top:0.1665in; text-align:left ! important; font-family:Liberation Sans;
writing-mode:page; margin-left:0in; margin-right:0in; text-indent:0in; }
        .P11 { font-size:18pt; margin-bottom:0.0835in; margin-top:0.0417in;
text-align:left ! important; font-family:Liberation Sans; writing-mode:page;
margin-left:0in; margin-right:0in; text-indent:0in; }
        .P2 { font-size:12pt; line-height:120%; margin-bottom:0.0972in;
margin-top:0in; font-family:Bitstream Charter; writing-mode:page; }
        .P3 { font-size:12pt; line-height:120%; margin-bottom:0.0972in;
margin-top:0in; font-family:Liberation Serif; writing-mode:page; margin-left:0in;
margin-right:0in; text-indent:0in; }
        .P4 { font-size:12pt; line-height:120%; margin-bottom:0.0972in;
margin-top:0in; font-family:Bitstream Charter; writing-mode:page; margin-left:0in;
margin-right:0in; text-indent:0in; }
        .P5 { font-size:12pt; line-height:120%; margin-bottom:0.0972in;
margin-top:0in; font-family:FreeSerif; writing-mode:page; margin-left:0in;
margin-right:0in; text-indent:0in; }
        .P6 { font-size:12pt; line-height:120%; margin-bottom:0in; margin-top:0in;
font-family:FreeSerif; writing-mode:page; margin-left:0in; margin-right:0in;
text-indent:0in; }
        .P7 { font-size:12pt; line-height:120%; margin-bottom:0.0972in;
margin-top:0in; font-family:Bitstream Charter; writing-mode:page; margin-left:0in;
margin-right:0in; text-indent:0.5in; }
        .P8 { font-size:12pt; line-height:120%; margin-bottom:0.0972in;
margin-top:0in; font-family:Liberation Serif; writing-mode:page; margin-left:0in;
margin-right:0in; text-indent:0.1252in; }
        .P9 { font-size:10.5pt; line-height:120%; margin-bottom:0.0972in;
margin-top:0in; font-family:Open Sans, Arial, sans-serif; writing-mode:page;
margin-left:0in; margin-right:0in; text-indent:0.3752in; color:#000000;
letter-spacing:normal; font-style:normal; font-weight:normal; }
        .Bullet_20_Symbols { font-family:OpenSymbol; }
        .T1 { font-weight:bold; }
        .T2 { font-style:italic; }
        .T3 { font-family:Bitstream Charter; }
        .T4 { font-family:Bitstream Charter; font-size:11pt; }
        <!-- ODF styles with no properties representable as CSS -->
        .Numbering_20_Symbols  { }
```

FIG. 4B

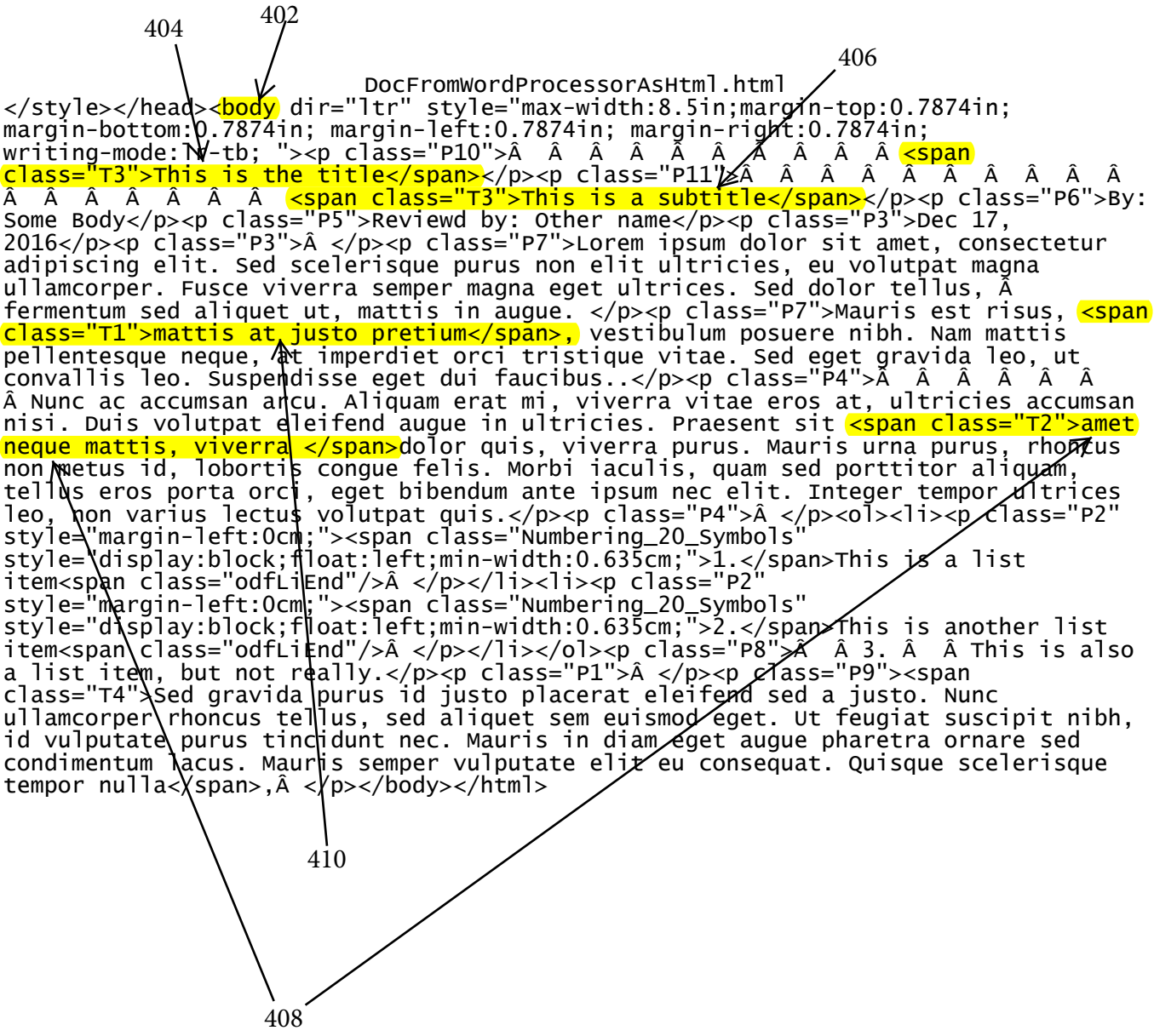
404
402
406
DocFromWordProcessorAsHtml.html
</style></head><body dir="ltr" style="max-width:8.5in;margin-top:0.7874in; margin-bottom:0.7874in; margin-left:0.7874in; margin-right:0.7874in; writing-mode:lr-tb; "> Â Â Â Â Â Â Â Â Â Â This is the title Â This is a subtitle By: Some Body Reviewd by: Other name Dec 17, 2016 Â Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed scelerisque purus non elit ultricies, eu volutpat magna ullamcorper. Fusce viverra semper magna eget ultrices. Sed dolor tellus, Â fermentum sed aliquet ut, mattis in augue. Mauris est risus, mattis at justo pretium, vestibulum posuere nibh. Nam mattis pellentesque neque, at imperdiet orci tristique vitae. Sed eget gravida leo, ut convallis leo. Suspendisse eget dui faucibus.. Â Â Â Â Â Â Â Nunc ac accumsan arcu. Aliquam erat mi, viverra vitae eros at, ultricies accumsan nisi. Duis volutpat eleifend augue in ultricies. Praesent sit amet neque mattis, viverra dolor quis, viverra purus. Mauris urna purus, rhoncus non metus id, lobortis congue felis. Morbi iaculis, quam sed porttitor aliquam, tellus eros porta orci, eget bibendum ante ipsum nec elit. Integer tempor ultrices leo, non varius lectus volutpat quis. Â 1.This is a list itemÂ 2.This is another list itemÂ Â Â 3. Â Â This is also a list item, but not really. Â Sed gravida purus id justo placerat eleifend sed a justo. Nunc ullamcorper rhoncus tellus, sed aliquet sem euismod eget. Ut feugiat suscipit nibh, id vulputate purus tincidunt nec. Mauris in diam eget augue pharetra ornare sed condimentum lacus. Mauris semper vulputate elit eu consequat. Quisque scelerisque tempor nulla,Â </body></html>
410
408

FIG. 4C

cleaned.html

```
<?xml version="1.0" encoding="UTF-8"?>
<html>
  <head>
    <link rel="stylesheet" type="text/css" href="styleguide.css">
  </head>
  <body>
    <p class="title">This is the title</p>
    <p class="subtitle">This is a subtitle</p>
    <p class="metadata">By: Some Body</p>
    <p class="metadata">Reviewd by: Other name</p>
    <p class="metadata">Dec 17, 2016</p>
    <p>Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed scelerisque
purus non elit ultricies, eu volutpat magna ullamcorper. Fusce viverra semper magna
eget ultrices. Sed dolor tellus, fermentum sed aliquet ut, mattis in augue.</p>
    <p>Mauris est risus, <span class="bold">mattis at justo pretium</span>,
vestibulum posuere nibh. Nam mattis pellentesque neque, at imperdiet orci tristique
vitae. Sed eget gravida leo, ut convallis leo. Suspendisse eget dui faucibus..</p>
    <p>Nunc ac accumsan arcu. Aliquam erat mi, viverra vitae eros at, ultricies
accumsan nisi. Duis volutpat eleifend augue in ultricies. Praesent sit <span
class="italic">amet neque mattis, viverra</span> dolor quis, viverra purus. Mauris
urna purus, rhoncus non metus id, lobortis congue felis. Morbi iaculis, quam sed
porttitor aliquam, tellus eros porta orci, eget bibendum ante ipsum nec elit.
Integer tempor ultrices leo, non varius lectus volutpat quis.</p>
    <ol>
      <li>This is a list item</li>
      <li>This is another list item</li>
      <li>This is also a list item, but not really.</li>
    </ol>
    <p>Sed gravida purus id justo placerat eleifend sed a justo. Nunc ullamcorper
rhoncus tellus, sed aliquet sem euismod eget. Ut feugiat suscipit nibh, id vulputate
purus tincidunt nec. Mauris in diam eget augue pharetra ornare sed condimentum
lacus. Mauris semper vulputate elit eu consequat. Quisque scelerisque tempor
nulla,</p>
  </body>
</html>
```

502 504 506 508 510

FIG. 5