I'm working on a paper about how we <u>use styles</u> in the <u>ICE</u> system to structure and format documents for print and the web. To help define the problem and show why an ICE-like approach is needed I decided to see how far I got trying to format a paper for print and the web using out-of-the-box word processing tools.
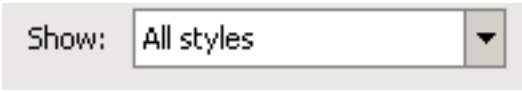
I covered this a little before when I <u>got cranky</u> about the HTML export in OpenOffice.org, but I haven't eviscerated Microsoft Word's HTML export here in detail.

So, here's the challenge.

Can I take my recent paper, <u>An integrated approach to preparing, publishing, presenting and preserving theses</u> for the ETD conference. I <u>wrote recently</u> about how I prepared this paper using ICE, meaning that of course I can produce both print and web versions, but how would I go using Word 2003 on Windows XP? Could I produce XHTML? This is, after all the year 2007 and the World Wide Web is in its late teens. (I didn't use Word 2007 because I don't have a copy, but I'll find one and try that too)

The biggest challenge, I knew, would be getting stuff that looks OK on screen to export to the correct HTML elements.

So here's the procedure: (I know this looks boring, but if you know a `<ul>` from a `<p>` you'll find it hysterically funny I promise)

1. Open a new blank document.

2. Look in the styles pane showing at the left of the page; only a couple of styles there so;

3. Select All styles from the Show drop down:

4. That's better – there are lots of heading styles and some HTML styles that should come in handy.

5. Select the `Title` style and type the title

6. Select the `Heading 1` style and type the first heading (lets forget about the abstract for now)

7. Type a few paragraphs of plain text.

8. Look in the style pane and work out that there's no blockquote style. Weirdly there is a built-in style called `HTML Preformatted`.

9. Save the document as a web page and note that the headings have worked.

10. Realize that you've saved it as a `.mht` whatever that is.

11. Save the document as HTML.

12. Look at the document in Firefox. Looks OK.

13. View source. Yuck!

14. Save the document as Filtered HTML . Hmmmm Filtered.

15. Have a look at the first bit of blockquote:

```
<p style='margin-top:13.05pt;margin-right:0cm;margin-
bottom:0cm;margin-left:

0cm;margin-bottom:.0001pt'><span lang=EN-US style='font-
```

```
size:11.0pt'>ICE-RS is

supported by the Systemic Infrastructure Initiative as part of
the Australian

Commonwealth Government's Backing Australia's Ability - An
Innovative Action

Plan for the Future
(http://backingaus.innovation.gov.au).</span></p>
```

Oh dear!

(Those extra spans with 11pt text are because I copied an pasted rather than typing: I'd need to turn that into default paragraph font. Easy enough on Windows, but apparently impossible in the Mac version of Word – they left that bit out!)

16. Use your 1337 skills to change the HTML that word exported to be proper: XHTML:

```
<blockquote>ICE-RS is

supported by the Systemic Infrastructure Initiative as part of
the Australian

Commonwealth Government's Backing Australia's Ability - An
Innovative Action

Plan for the Future
(http://backingaus.innovation.gov.au).</blockquote>
```

17. Load the document back in to Word to see what happens.

18. Observe that Word has turned the blockquote abck into a paragraph of Normal text.

19. Save the document.

20. Look at the source.

```
<blockquote style="margin-top: 5pt; margin-bottom: 5pt;">

<p class="MsoNormal">ICE-RS is supported by the Systemic
Infrastructure

Initiative as part of the Australian Commonwealth Government's
Backing

Australia's Ability - An Innovative Action Plan for the Future

(http://backingaus.innovation.gov.au).</p>

</blockquote>
```

Wow! There's still a blockquote there, but guess what? In Word you can't *see* that it's there!

21. Add some more text to the blockquote by going to the end of the paragraph.

22. Save the document.

23. Check that yes, you have added to the blockquote.

24. Back in the document, experiment with the demote button, the left-facing arrow **<–**. The

first part of the blockquote won't move, but the second will. Here it is as rendered on-screen in Word.

derived from the successful proposal document is available on the ICE website (Sefton 2006b).

> ICE-RS is supported by the Systemic Infrastructure Initiative as part of the Australian Commonwealth Government's Backing Australia's Ability – An Innovative Action Plan for the Future (http://backingaus.innovation.gov.au).

Test adding more text.

25. Guess what the HTML for that might look like. Here it is rendered in Firefox:

documents, collaborate, manage, publish and deposit their work in repositories. An overview of the project, derived

> ICE-RS is supported by the Systemic Infrastructure Initiative as part of the Australian Commonwe:
> Future (http://backingaus.innovation.gov.au).

The project proposal describes the key aim; to provide more flexible content than having all documents delivered in

26. Look at source to see where the text 'Test adding more text ' has gone.

```
<blockquote style='margin-top:5.0pt;margin-bottom:5.0pt'>

<p class=MsoNormal>ICE-RS is supported by the Systemic
Infrastructure

Initiative as part of the Australian Commonwealth Government's
Backing

Australia's Ability – An Innovative Action Plan for the Future
(<a

href="http://backingaus.innovation.gov.au/">http://backingaus.i
nnovation.gov.au</a>).</p>

<p class=MsoNormal style='margin-left:-180.0pt'>Test adding
more text.</p>

</blockquote>
```

You see that? Word has helpfully put on margin-left:-180.0pt. Hmm a left margin of minus one hundred and eighty points.

27. Give up in disgust. I can't see a way to get Word 2007 to make a blockquote.

(And I tried a couple of other things too, like guessing that if I used a style like `HTML Blockquote` Word might magically Do (Nearly) The Right Thing the way it does with `HTML Preformatted` style. It doesn't it makes a paragraph with class `HTMLBlockquote` but with the wrong CSS. Oh well.)

Actually I can't give up because I have yet to play with the lists.

There's a bit in my paper where I have a blockquote with a list embedded in it. That's pefectly possible in ICE, but would be very hard in Word. So lets look for an easier case.

So how would we do this example with a numbered list, with some pre-format ted text embedded in it? (note that the actual document has a rendering mistake in it that I have fixed here).

1. Initially the handle will resolve to the server-side ICE repository, which because it is in the Subversion system is web-addressable, although usually authentication will be required.

   The author need not worry about handles at all: they can use links in the usual way to manage their content and the system will manage the creation and management of handles when content is exported from the system.

   An internal link would look like an ICE URL:

   ```
   http://localhost:8000/some-path
   ```

   When exported it would use a handle resolver:

   ```
   http://myresolver.edu.au/hdl/4435387897435
   ```

   Which would resolve to the ICE repository:

   ```
   http://myrepository.myuni.edu.au/some-path
   ```

2. When the ETD goes from authoring to the examination phase the software will export it from the ICE site to a web site, media such as a CD-ROM or access controlled repository: the handles will then resolve to any online version.

I used the built-in styles `List Number 1`, `List Continue` and `HTML Preformatted` (with a hit on the indent button).

Result *looks* OK in Firefox but it's a million miles from being XHTML. The list is not a list at all, it's a paragraph with some creative formatting and a bunch of non-breaking spaces. Really. I used to know this stuff intimately – see my article on geeting XML out of Word for xml.com.

```
<p class=MsoListNumber><span lang=EN-US>1.<span style='font:7.0pt
"Times New Roman"'>     

</span></span><span lang=EN-US>Initially the handle will resolve
to the

server-side ICE repository, which because it is in the Subversion
system is

web-addressable, although usually authentication will be
required.</span></p>
```

At this point I'd give up, but I love them lists too much. So, why not try ignoring styles and using the formatting buttons.

Nope. Still no list. It's the same nonsense but without any CSS.

```
<p style="margin-left: 36pt; text-indent: -18pt;">

<span lang="EN-US">1.<span style="font-family: &quot;Times New
Roman&quot;;

font-style: normal; font-variant: normal; font-weight: normal;

 font-size: 7pt; line-height: normal; font-size-adjust: none;
```

```
font-stretch: normal;">      </span>

</span><span lang="EN-US">Initially the handle will resolve to the
server-side ICE repository,



which because it is in the Subversion system is web-addressable,
although

usually authentication will be required.</span></p>
```

Conclusion

This sux.

The worst bit was when I managed to get a word document that contained a blockquote that is invisible through the editing interface, but which creates nightmares like invisible paragraphs with their left margin miles off the screen.

If you gave Word 2003 to somebody and asked them to write a paper that could be given to a fussy HTML publisher and also printed with nice headers and footers, or saved to PDF then they'd be stuck.

Which I kind of knew, which is why we invented ICE. But I needed to go through this so I can show the results for the paper I'm writing.

Next up, OpenOffice.org Writer. What do you think OOo fans? Will it do any better? An how about Google Docs?