

Legacy HTML Support in PDFReactor

Differences between HTML 4.0 and XHTML 1.0

1. Tags must be closed

```
HTML 4.0: <p>This p element isn't closed  
XHTML 1.0: <p>This p element is closed</p>
```

2. Attributes must have values which are enclosed by doublequotes

```
HTML 4.0: <input type=text value=test>  
XHTML 1.0: <input type="text" value="test">
```

3. Tag names must be lower case

```
HTML 4.0: <P>This is a Paragraph</P>  
XHTML 1.0: <p>This is a paragraph</p>
```

4. The document must be wellformed

```
HTML 4.0:<p><strong>This should be strong</p></strong>  
XHTML 1.0:<p><strong>This should be strong</strong></p>
```

The cleanup process

The cleanup capability provides 2 different cleanup tools:

1. **Cyberneko HTML parser**

This HTML parser fixes the following XHTML compatibilities:

- adds missing parent elements
- automatically closes elements
- handles mismatched end tags

2. **jTidy**

JTidy is a Java port of HTML Tidy, a HTML syntax checker and pretty printer.

jTidy provides the following features (among others):

- Missing or mismatched end tags are detected and corrected
- End tags in the wrong order are corrected
- Recovers from mixed up tags
- Adding the missing "/" in end tags for anchors
- Perfecting lists by putting in tags missed out
- Missing quotes around attribute values are added
- Unknown/Proprietary attributes are reported
- Proprietary elements are recognized and reported as such
- Tags lacking a terminating '>' are spotted

Some content to cleanup

Here follows some non XHTML compliant content, in order to see the cleanup process in action. Please open the "Advanced tab" for more information about the cleaned document.

This p element isn't closed.

This text is enclosed in uppercase p elements.

These elements are incorrectly nested.