Hyphenation rules the nation

Thursday 23 July 2009 04:17

As I mentioned earlier, I'm working on a Perl module called HTML::Hyphenate, and now rolling out a test on this blog. I've added a filter in my main Bricolage web template that inserts all the soft hyphens in the body of every page. So if your browser is not totally retarded you could see some words on this page being broken onto more than one line, and the text of the posts should be nicely justified. If your browser is somewhat retarded you might just see all the inserted soft hyphens as dashes and read this text very slow. My sort of english doesn't have very long words, but since I set the hyphenation algorithm to use a minimum word length of five characters words like "water" can already be broken into "wa−ter". Languages other than en_US should also be handled correctly, and even scripts like Sanskrit and Mongolian should just work. And of course Arabic, because that doesn't have hyphenation.

Different browsers do of course different things. Opera for example places the dashes as which the soft hyphens are revealed as at breaks like they appear outside of the text box. Which is ugly. Safari does it better, lynx seems to ignore the soft hyphens, Amaya shows them as dashes. Browsers could of course implement the algorithms themselves and do the hyphenation without the author supplying all the points, but doing that suddenly won't make the web look much better. As soon as you start with hyphenation a lot of things can break. You need to specify the language for every word so for example german words don't get weird breaks as if they were english, and insert nobr tags or CSS equivalents to prevent ugly stuff, like "Star−ck". So now I have to look into CSS to see if that in some way can be used to tell that words should be broken differently if they are the last word in a sentence or paragraph.

by Roland van Ipenburg
http://www.xs4all.nl/~ipenburg/blog/posts/work/2009/07/23/hyphenation−rules−the−nation/