HTMLGenerator

This page was created mainly to answer the *How do I handle bad HTML content* FAQ.

Cocoon uses JTidy

To allow HTML documents to be used as input to pipelines, Cocoon uses JTidy as the basis of its HTMLGenerator, to parse HTML, allowing many less-than-perfect HTML documents to be converted to XML (technically SAX events).

Handling even more lousy HTML

Before release 2.0.4, the HTMLGenerator did not allow all JTidy options to be set.

From 2.0.4, the "jtidy-config" configuration element of the HTMLGenerator points to a properties file that can be used to set all JTidy options, giving better control on the processing of HTML input (thanks Sylvain Wallez!).

References

- HTMLGenerator:[http://xml.apache.org/cocoon/userdocs/generators/html-generator.html official documentation].
- cocoon-users mailing list, Handling lousy HTML, September 2002.

Notes and comments

- The CyberNeko HTML Parser could be used instead of JTidy, maybe giving the choice between the two would help?
- See also HTMLTidy