Website Input - Bug #766

Page encoding is not determined correctly

07/13/2014 04:40 AM - Luke Murphey

Status:	Closed	**Start date:**	07/13/2014
Priority:	Normal	**Due date:**	
Assignee:	Luke Murphey	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:	0.8		

Description

History

#1 - 07/13/2014 04:40 AM - Luke Murphey

- Target version set to 0.8

#2 - 07/13/2014 04:41 AM - Luke Murphey

http://answers.splunk.com/answers/144939/website-input-charset-problem

#3 - 07/13/2014 04:46 AM - Luke Murphey

This is how html5lib does it:

"If no encoding is specified, the parser will attempt to detect the encoding from a <meta> element in the first 512 bytes of the document (this is only a partial implementation of the current HTML 5 specification).

If no encoding can be found and the chardet library is available, an attempt will be made to sniff the encoding from the byte pattern.

If all else fails, the default encoding will be used. This is usually Windows-1252, which is a common fallback used by Web browsers."

#4 - 07/13/2014 04:50 AM - Luke Murphey

- Status changed from New to In Progress

#5 - 07/13/2014 06:11 AM - Luke Murphey

- Status changed from In Progress to Closed

- % Done changed from 0 to 100