Metacat - Bug #1538

Entity/Character Refeference Conversion Problems

04/28/2004 09:29 AM - Dan Higgins

Status:	Resolved	**Start date:**	04/28/2004
Priority:	Immediate	**Due date:**	
Assignee:	Saurabh Garg	**% Done:**	0%
Category:	metacat	**Estimated time:**	0.00 hour
Target version:	1.5	**Spent time:**	0.00 hour
Bugzilla-Id:	1538		

Description

In Morpho, we have run into some problem with the use of special, non-ascii characters (like the 'degree' symbol or greek 'mu'). [Any character represented by a byte with a decimal value > 127 is in this class of special characters.] These characters have been copied from Word or PDF documents into Morpho fields and then put into eml xml docs. Unfortunately, they are not not necessarily in the correct format for xml documents and have caused parser problems.

The solution that was implemented in Morpho was to use entity/character references. Any character with a value greater than 127 is written as 'xx;' where 'xxx' is the decimal value of the character. On a Windows machine, the 'deg' symbol becomes '°' and 'mu' becomes 'µ'. XML parsers automatically convert these character entities to the character for display, but the conversion depends on the assumed character set.

The metacat problem is that when one submits a document containing such character references (xx;) and then reads the document back, one does not get the character refenence, but rather the character itself! I assume this is due to the XML parser. This is a violation of the idea that metacat should return exactly the same data given it.

Morpho already handles this by converting back to character references any info sent it by Metacat with character values greater than 127. But metacat actually sends back the wrong character for some symbols! (e.g. a 'mu' becomes a '1/4' symbol. I assume this is due to different character set assumption under linux and windows. In any case, there is some data corruption here that we should figureout how to avoid.

History

#1 - 12/15/2004 12:58 PM - Matt Jones

Its not clear to me that we are handling non-ascii characters properly. Internet Explorer and some other applications reject several documents that are in metacat as invalid because of these non-ascii characters. We need to determine exactly what the proper behavior is. We probably need to change the xml declaration to indicate the proper character set encoding if non-ascii characters are present. Andrea and Veronique reported several similar issues with character problems in the most recent version of morpho (1.5.1).

#2 - 12/20/2004 03:26 PM - Saurabh Garg

Changed the code of the normalize function in MetaCatUtil.java. Earlier code was not taking care of characters above 123.
In DBSAXHandler.java, added call to normalize function before text is written to db.

So now the data is normalized before it is stored in the DB. When the dp is read again, the XX; form is returned. It can be denormalized, if that is considered appropriate.

Will close this bug after some more testing.

#3 - 01/13/2005 05:31 PM - Saurabh Garg

Closing the bug.

#4 - 03/27/2013 02:17 PM - Redmine Admin

Original Bugzilla ID was 1538