Metacat - Bug #3815

Ampersand character not correctly encoded

02/09/2009 05:35 PM - Shaun Walbridge

Status:	Resolved	**Start date:**	02/09/2009
Priority:	Normal	**Due date:**	
Assignee:	ben leinfelder	**% Done:**	0%
Category:	metacat	**Estimated time:**	0.00 hour
Target version:	2.0.0	**Spent time:**	0.00 hour
Bugzilla-Id:	3815		

Description

Ampersands are encoded as "&" within register-dataset.cgi in normalize(), but documents uploaded have a "%26amp;" entry instead. "%26" is the urlencoded version of "&" and 0x0026 is the Unicode code-point.

An example document exhibiting the behavior:
http://knb.ecoinformatics.org/knb/metacat?action=read&qformat=nceas&docid=nceas.912.8

The organization is set to "U.S. Fish %26amp; Wildlife Service".

History

#1 - 02/10/2009 12:33 PM - Shaun Walbridge

To properly fix this bug, we'll need to add tests for character conversions, to make sure that we aren't introducing regressions with our fixes. Add test documents which break the system, and then afterward add the necessary fixes.

register-dataset.cgi has no testing, and may also be the source of this bug.

#2 - 02/10/2009 01:05 PM - Matt Jones

Register-dataset.cgi may not have any testing itself, but it is based on Metacat.pm which does have a test suite in the module definition library. So, I think that extending the tests in Metacat.pm should help in covering multiple scripts like register-dataset.cgi that might make use of it for inserting and updating data to metacat.

#3 - 10/27/2011 08:23 PM - ben leinfelder

Using the dev skin to load an XML document with an ampersand encoded as: & kept the character intact. To me, this indicates that Metacat's servlet API is correctly handling the character. If the register-dataset.cgi or Metacat.pm is doing something to encode this additionally, that might account for the double encoding.

#4 - 10/27/2011 10:23 PM - ben leinfelder

Using the Java MetacatClient API to insert a document with & also worked fine (no additional encoding of the & symbol).

#5 - 10/28/2011 01:11 PM - Shaun Walbridge

The specific example looks to have its origins in delNormalize (sic) within register-dataset.cgi -- the first regex operator replaces '&' with '&' but the last regex operator then replaces '&' with '%26'. I'd recommend removing these functions and using existing modules for encoding/decoding of the XML. There are a couple of options [1], though perhaps just fixing the symptom is good enough for the time being.

1. http://stackoverflow.com/questions/1137790/how-can-i-escape-text-for-an-xml-document-in-perl

#6 - 10/28/2011 01:23 PM - ben leinfelder

I only see delNormalize() being called from the deleteData() function in register-dataset.cgi -- but perhaps similar code is lurking somewhere else? Is this something you (Shaun) can look into?

#7 - 11/09/2011 02:28 PM - ben leinfelder

I just tried the registry on a test machine running most recent Metacat trunk and the ampersand was correctly encoded for XML (and only once):

<title>Testing & Stuff</title>

#8 - 03/27/2013 02:24 PM - Redmine Admin

Original Bugzilla ID was 3815