Infrastructure - Task #8121

Fix MIME type and formatId for KML files

2017-06-30 19:37 - Matthew Jones

Status:	Closed	**Start date:**	2017-06-30
Priority:	High	**Due date:**	
Assignee:	Jing Tao	**% Done:**	100%
Category:	dataone-cn-metacat	**Estimated time:**	0.00 hour
Target version:	CCI-2.3.5		
Milestone:	None	**Story Points:**	
Product Version:	*		

Description

The Arctic Data Center data team is trying to upload some KML files, and noticed an oddity in the formatId and mime type for KML files. The proper MIME type should be application/vnd.google-earth.kml+xml (see https://developers.google.com/kml/documentation/kml_tut#kml_server), but the DataONE formatId and mime type are set to application/vnd.google-earth.kml xml, where a space was substituted for the + sign.

At a minimum, the MIME type should be corrected in our formats list (https://cn.dataone.org/cn/v2/formats), but I think the formatId itself should also be corrected. It appears there are only five objects using this formatId in production: https://cn.dataone.org/cn/v2/object?formatId=application/vnd.google-earth.kml%20xml , so I think we could change the formatId in our vocabulary, and work with the owners of those 5 objects to update their formatId. Thoughts?

Related issues:

Related to Infrastructure - Task #8128: Correct invalid formatIDs in production	**Closed**	**2017-07-11**
Blocked by Infrastructure - Bug #8122: Metacat is double-decoding incoming ur...	**Closed**	**2017-06-30**

History

#1 - 2017-06-30 20:11 - Dave Vieglais

- Target version set to CCI-2.3.6

- Priority changed from Normal to High

- Assignee set to Jing Tao

This appears to be a bug in Metacat since the source of the formatIDs is correct:

https://repository.dataone.org/software/cicore/trunk/cn-buildout/dataone-cn-metacat/usr/share/metacat/debian/objectFormatListV2.xml

Notes on adding formatIds:
http://jenkins-1.dataone.org/jenkins/job/DataONE-Operations-Manual/ws/operations/_build/html/object_format_registration/manually_adding_object_formats.html

#2 - 2017-06-30 21:28 - Dave Vieglais

Tested the upload / update script using an echo server. The uploaded content is available as expected on the server side.

From Rob via slack:

(space)

original pathInfo: /formats/application/rdf xml
original requestURI: /metacat/d1/cn/v2/formats/application/rdf%20xml
new pathinfo: /formats/application/rdf%20xml
After decoded: application/rdf xml

(plus)

original pathInfo: /formats/application/rdf+xml
original requestURI: /metacat/d1/cn/v2/formats/application/rdf+xml
new pathinfo: /formats/application/rdf+xml
After decoded: application/rdf xml

So it appears the issue lies in either tomcat passing on improperly escaped content, or metacat improperly unescaping content being received.

#3 - 2017-07-01 10:35 - Dave Vieglais

- Status changed from New to In Progress

- % Done changed from 0 to 30

On further consideration, the fundamental issue is that the script:

insertOrUpdateObjectFormatList.sh

is using the "-d" switch with curl to post form data.

This mechanism expects the data to be application/x-www-form-urlencoded by the submitter. Curl itself does not urlencode data unless the --data-urlencode switch is provided.
The result is the plain XML document being sent as form data, the server is urldecoding when being processed, and the "+" characters are thus being faithfully converted to spaces.

The fix is to replace the "-d" switches in the script with "--data-urlencode" unless the data being sent has already been specifically urlencoded.

Then upload the formatId document and verify correct content is available on the server.

#4 - 2017-07-01 10:46 - Dave Vieglais

- Target version changed from CCI-2.3.6 to CCI-2.3.5

After manually correcting the script on cn-dev-unm-1 and executing, the resulting format list from the CN appears correct:

curl -s "https://cn-dev-unm-1.test.dataone.org/cn/v2/formats" | grep svg
image/svg+xml

svg

#5 - 2017-07-01 11:06 - Dave Vieglais

IMPORTANT:

Before deploying this fix to all CNs, issue #8122 must be corrected, otherwise it is not possible to retrieve format information for any formatIds that have a "+" in them. For example, requesting @application/rdf+xml@ from cn-dev-unm-1:

curl "https://cn-dev-unm-1.test.dataone.org/cn/v2/formats/application%2Frdf%2Bxml"
<?xml version="1.0" encoding="UTF-8"?>
The format specified by application/rdf xml does not exist at this node.

but that format is present:

curl -s "https://cn-dev-unm-1.test.dataone.org/cn/v2/formats" | xml sel -t -m "//formatId[text()='application/rdf+xml']" -c ..

application/rdf+xml
Resource Description Framework
DATA

rdf

#6 - 2017-07-01 11:07 - Dave Vieglais

- Blocked by Bug #8122: Metacat is double-decoding incoming urls on the CNs added

#7 - 2017-07-06 16:00 - Jing Tao

After we upgraded Metacat, this link works now:
https://cn-dev.test.dataone.org/cn/v2/formats/application/rdf%2Bxml

#8 - 2017-07-11 18:12 - Jing Tao

- Category set to dataone-cn-metacat

#9 - 2017-07-11 19:40 - Dave Vieglais

- Related to Task #8128: Correct invalid formatIDs in production added

#10 - 2017-07-17 22:52 - Jing Tao

- Status changed from In Progress to Closed

- % Done changed from 30 to 100