CN REST - Bug #7879

CN fails to correctly decode PID with encoded character

2016-09-06 19:06 - Mark Servilla

Status:	Closed	**Start date:**	2016-09-06
Priority:	High	**Due date:**	
Assignee:	Jing Tao	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:			
Story Points:			

Description

The production CN fails to resolve a PID that contains an encoded character, specifically the space " " character when encoded as "%20". The TDAR MN has successfully synchronized objects with PIDs like "doi:10.6067:XCV8639NVH_format=d1rem2010-11-05%2014:10:21.616". When attempting to read either the system metadata or object using curl, the CN returns a 404 not found error and the PID in the exception message is not the PID requested -- the "%20" character between the date and time strings is replaced with an actual space " ":

curl -s -X GET "https://cn.dataone.org/cn/v2/object/doi:10.6067:XCV8639NVH_format=d1rem2010-11-05%2014:10:21.616"
<?xml version="1.0" encoding="UTF-8"?>
No system metadata could be found for given PID: doi:10.6067:XCV8639NVH_format=d1rem2010-11-05 14:10:21.616

Note that the correct PID is returned in listObjects and is correctly entered into the Metacat system metadata and identifier tables.

History

#1 - 2016-09-06 19:58 - Jing Tao

I looked at the catalina.out file and found:

CN Dispatching: /d1/cn/v2/object/doi:10.6067:XCV8639NVH_format=d1rem2010-11-05%2014:10:21.616
original pathInfo: /object/doi:10.6067:XCV8639NVH_format=d1rem2010-11-05%2014:10:21.616
original requestURI: /metacat/d1/cn/v2/object/doi:10.6067:XCV8639NVH_format=d1rem2010-11-05%2014:10:21.616
new pathinfo: /object/doi:10.6067:XCV8639NVH_format=d1rem2010-11-05%2014:10:21.616
After decoded: doi:10.6067:XCV8639NVH_format=d1rem2010-11-05 14:10:21.616
org.dataone.service.exceptions.NotFound: No system metadata could be found for given PID: doi:10.6067:XCV8639NVH_format=d1rem2010-11-05 14:10:21.616

I can tell the cn forwarded the url to Metacat without decoding the identifier but Metacat decoded the object identifier.

I also looked at the code and found Metacat has the code to decode the identifier when users try to get the objects or read the systemmetadata.

I can understand this - sometimes the identifier has special characters and it has to be encoded in order to go through the rest url. In order to get the original identifier, we need to decode it.

But this case, we don't need to decode it. However, Metacat is not smart enough :(

In what scenario, we need to decode the identifier or don't need? (This case seems to me it uses a space in their identifier which is illegal in DataONE federation.)

#2 - 2016-09-06 22:09 - Dave Vieglais

- Status changed from New to In Progress

- % Done changed from 0 to 30

This appears to be working OK for me.

The identifier:
https://cn.dataone.org/cn/v2/object/doi:10.6067:XCV8639NVH_format=d1rem2010-11-05%2014:10:21.616

Should be URL encoded as:

https://cn.dataone.org/cn/v2/object/doi:10.6067:XCV8639NVH_format=d1rem2010-11-05%252014:10:21.616

When used in resolve:

curl "https://cn.dataone.org/cn/v2/resolve/doi:10.6067:XCV8639NVH_format=d1rem2010-11-05%252014:10:21.616"

or getSystemMetadata:

curl "https://cn.dataone.org/cn/v2/meta/doi:10.6067:XCV8639NVH_format=d1rem2010-11-05%252014:10:21.616"

it appears to work just fine. I suspect it was an encoding issue on the client side that caused the problem, though during the call this morning it seemed that the error was confirmed.

#3 - 2016-09-07 02:26 - Matthew Jones

Our docs disallow spaces and other non printing characters in Identifiers. See the identifiers docs, which say:

"Structure
The characters that may appear in an identifier string acceptable to the DataONE system is constrained by the XMLSchema definition (Types.Identifier), which is essentially a string of length greater than zero but less than 800 characters with no whitespace (spaces, tabs, non-printing characters, carriage returns, new lines)."

#4 - 2017-03-28 16:41 - Dave Vieglais

- Status changed from In Progress to Closed

- % Done changed from 30 to 100

Matthew Jones wrote:

> Our docs disallow spaces and other non printing characters in Identifiers. See the identifiers docs, which say:
>
> "Structure
> The characters that may appear in an identifier string acceptable to the DataONE system is constrained by the XMLSchema definition (Types.Identifier), which is essentially a string of length greater than zero but less than 800 characters with no whitespace (spaces, tabs, non-printing characters, carriage returns, new lines)."

The identifier is "doi:10.6067:XCV8639NVH_format=d1rem2010-11-05%2014:10:21.616" thus there is no space in the identifier string as it appears to DataONE.

DataONE operations appear to work as expected when the identifier is properly URL encoded.