Infrastructure - Bug #6391

Science metadata files with different checksums on CN and MN - encoding

2014-09-18 16:17 - Skye Roseboom

Status:	Closed	**Start date:**	2014-11-04
Priority:	Normal	**Due date:**	2015-01-06
Assignee:	Jing Tao	**% Done:**	100%
Category:	Metacat	**Estimated time:**	0.00 hour
Target version:	CCI-2.0.0		
Milestone:	None	**Story Points:**	
Product Version:	*		

Description

Discovered that some science metadata docs on MN KNB and the CN have different checksums. Further investigation shows what appears to be encoding differences which lead to the different checksums.

Examples:
(title element)
https://cn.dataone.org/cn/v1/object/doi:10.5063/AA/knb.277.1
https://knb.ecoinformatics.org/knb/d1/mn/object/doi:10.5063/AA/knb.277.1

```
    (method section)
https://cn.dataone.org/cn/v1/object/doi:10.5063/AA/nceas.985.1
https://knb.ecoinformatics.org/knb/d1/mn/object/doi:10.5063/AA/nceas.985.1
```

Related issues:

Related to Infrastructure - Story #3309: get() downloads have inconsistent mi...	**Closed**	**2012-10-08**	
Related to Infrastructure - Task #6568: Create or update a metadata object wi...	**Closed**	**2014-11-13**	**2014-11-14**
Related to Infrastructure - Task #7042: Create an element for the character s...	**Closed**	**2015-04-14**	

History

#1 - 2014-09-18 17:30 - Dave Vieglais

RFC 7303 Sections 8.8 and 8.9 are particularly relevant here. http://www.rfc-editor.org/rfc/rfc7303.txt

HTTP headers from the cn:

Content-Type: text/xml;charset=UTF-8

and from the mn:

Content-Type: text/xml

Both the MN and the CN are operating incorrectly, though the CN is worse.

The MN SHOULD be setting the charset parameter when transmitting over HTTP.

The CN output is an example of incorrect encoding of output. In that case, the charset parameter of the HTTP header specifies UTF-8, but the encoding parameter of the XML document indicates ISO-8859-1. HTTP stream readers will likely interpret the content as UTF-8, though if the file is saved without a BOM, then an editor will recognize the file as ISO-8859-1.

#2 - 2014-09-18 19:10 - Robert Waltz

In the web.xml of the DataONE_CN_Rest project there is a filter applied to every request named CharacterEncodingFilter that forces all encoding to be UTF-8. We can either take the filter off, or make its application more selective.

So, if this bug is still seen to apply to Metacat, then we should make a second bug report for cn_rest_service as well.

#3 - 2014-09-24 18:10 - Skye Roseboom

- Due date set to 2014-09-24

- Start date set to 2014-09-24

- Target version set to CCI-1.5.0

#4 - 2014-09-25 16:57 - Dave Vieglais

- Due date changed from 2014-09-24 to 2014-09-25

Conclusion after examining literature and the requirements of the get() operation:

- The server should follow the recommendations of RFC7303 and related operational specifications such as RFC6838 and RFC7231
- If the client is making an exact copy (e.g. during CN sync of metadata, and for replicas), the client MUST stream the exact bytes of the HTTP GET body without alteration
- The client SHOULD make accessible any MIME and other descriptive information provided by the origin
- The CNs SHOULD record the MIME type and other descriptive information about the content so that this information can be passed on to clients.

#5 - 2014-10-29 00:04 - Jing Tao

- Due date changed from 2014-09-25 to 2014-10-29

MN seems ok. Please see 8.3 http://www.rfc-editor.org/rfc/rfc7303.txt.

#6 - 2014-10-29 18:33 - Jing Tao

First step is to implement this:
If the client is making an exact copy (e.g. during CN sync of metadata, and for replicas), the client MUST stream the exact bytes of the HTTP GET body without alteration

#7 - 2014-10-29 20:09 - Jing Tao

in the read actions (such as get and getReplica), bytes are read directly from files. So there is no alteration.

However, in the save actions (such as create and update), there is transform from an input stream to a string by using default utf-8 encoding:

CNode.create(inputStream) ---> D1Node.create(inputStream) ---> D1Node.insertOrUpdateDocument(xmlString)--->MetacatHandler.handleInsertOrUpdateAction(xmlString). The inputStream is transformed to a string by use UTF-8 encoding.

MNode.create(inputStream) ---> D1Node.create(inputStream) --->

D1Node.insertOrUpdateDocument(xmlString)--->MetacatHandler.handleInsertOrUpdateAction(xmlString). The inputStream is transformed to a string by use UTF-8 encoding.

MNode.update(inputStream) ---> D1Node.insertOrUpdateDocument(xmlString)--->MetacatHandler.handleInsertOrUpdateAction(xmlString). The inputStream is transformed to a string by use UTF-8 encoding.

#8 - 2014-11-03 18:19 - Jing Tao

- Due date changed from 2014-10-29 to 2014-11-03

During the replication in Metacats, Metacat change the input stream from another host to a string using the default encoding, then save the string to the file. We need to save the input stream directly.

#9 - 2014-11-04 18:51 - Jing Tao

- Due date changed from 2014-11-03 to 2014-11-04

- Status changed from New to In Progress

#10 - 2014-11-13 18:38 - Jing Tao

- Status changed from In Progress to Testing

- Due date changed from 2014-11-04 to 2014-11-13

#11 - 2014-11-14 00:03 - Jing Tao

- Due date changed from 2014-11-13 to 2014-11-14

- Target version changed from CCI-1.5.0 to CCI-1.5.1

#12 - 2014-11-14 00:04 - Jing Tao

The features will be released in cci-1.5.0 is:

https://redmine.dataone.org/issues/6568

#13 - 2015-01-06 18:41 - Dave Vieglais

- Due date changed from 2014-11-14 to 2015-01-06

- Target version changed from CCI-1.5.1 to CCI-2.0.0

#14 - 2015-04-14 20:49 - Jing Tao

- Related to Task #7042: Create an element for the character set encoding in the system metadata schema added

#15 - 2015-04-14 20:50 - Jing Tao

- Status changed from Testing to Closed

- % Done changed from 0 to 100

I created a new ticket to generate an option element for the character set encoding in the system metadata.
https://redmine.dataone.org/issues/7042
close the ticket.