Infrastructure - Bug #7601

CN checksum inconsistencies

2016-01-21 20:09 - Ben Leinfelder

Status:	New	**Start date:**	2016-01-21
Priority:	High	**Due date:**	
Assignee:	Dave Vieglais	**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:			
Milestone:	None	**Story Points:**	
Product Version:	*		

Description

While transferring test data from production to the sandbox-2 environment I noticed failures for a group of pids.
I'll use an example to illustrate (doi_10.5066_F71C1TV7)
https://cn.dataone.org/cn/v2/meta/doi_10.5066_F71C1TV7
CN.SystemMetadata reports checksum as:

46178da6192263921eb755940d716725

Whereas calculating it from disk gives this:

MD5(/var/metacat/documents/autogen.2013062508395355978.1)= efc11787f789b45db29999fb4bd8d745

The byte size is also off.

16739

On disk:

-rw-r--r-- 1 tomcat7 tomcat7 16529 Jun 25 2013 /var/metacat/documents/autogen.2013062508395355978.1

There are ~70 similar pids that have issues (perhaps more) from our test corpus. They are from the now defunct USGS MN.

I'm not sure what our strategy is since the original MN is not online any longer so we cannot get the "original" bytes from that.

History

#1 - 2016-01-21 20:10 - Ben Leinfelder

Here are the pids that are similar

doi_10.5066_F7028PGW
doi_10.5066_F7028PHB
doi_10.5066_F71C1TV7
doi_10.5066_F71J97PQ
doi_10.5066_F71N7Z32
doi_10.5066_F7251G5C
doi_10.5066_F7319SV1
doi_10.5066_F73X84MM
doi_10.5066_F7416V18
doi_10.5066_F7445JD4
doi_10.5066_F74T6G97
doi_10.5066_F75M63MZ
doi_10.5066_F75M63ND
doi_10.5066_F75Q4T2R
doi_10.5066_F75T3HFM
doi_10.5066_F77942P6
doi_10.5066_F77P8WBZ
doi_10.5066_F78C9T8T
doi_10.5066_F78K771Z
doi_10.5066_F78W3B9C
doi_10.5066_F7959FHN
doi_10.5066_F79K485B
doi_10.5066_F7BZ6409

doi_10.5066_F7C8277K
doi_10.5066_F7CF9N17
doi_10.5066_F7CF9N2P
doi_10.5066_F7CJ8BFJ
doi_10.5066_F7DV1GTW
doi_10.5066_F7F18WPR
doi_10.5066_F7F47M21
doi_10.5066_F7FJ2DQ9
doi_10.5066_F7G15XT4
doi_10.5066_F7G44N7V
doi_10.5066_F7G73BN5
doi_10.5066_F7J9649N
doi_10.5066_F7JM27JM
doi_10.5066_F7K07262
doi_10.5066_F7K64G1Q
doi_10.5066_F7KP8035
doi_10.5066_F7KW5CZV
doi_10.5066_F7MS3QPX
doi_10.5066_F7MS3QQC
doi_10.5066_F7N877RP
doi_10.5066_F7NS0RVR
doi_10.5066_F7P26W3W
doi_10.5066_F7PC30CP
doi_10.5066_F7PK0D4F
doi_10.5066_F7Q23X6T
doi_10.5066_F7QJ7F88
doi_10.5066_F7QR4V2G
doi_10.5066_F7RF5S1S
doi_10.5066_F7RJ4GDN
doi_10.5066_F7S180GD
doi_10.5066_F7S46PZ7
doi_10.5066_F7SF2T6M
doi_10.5066_F7ST7MSF
doi_10.5066_F7TB14W6
doi_10.5066_F7TH8JND
doi_10.5066_F7TQ5ZHH
doi_10.5066_F7V122QQ
doi_10.5066_F7VM497Z
doi_10.5066_F7VX0DHC
doi_10.5066_F7WM1BBW
doi_10.5066_F7WW7FK9
doi_10.5066_F7X9288R
doi_10.5066_F7XK8CH5
doi_10.5066_F7Z31WN0

#2 - 2016-01-21 20:11 - Ben Leinfelder

- Description updated

#3 - 2016-01-21 20:39 - Ben Leinfelder

- Description updated

#4 - 2016-01-22 17:45 - Dave Vieglais

Looking at the first item in the list: doi_10.5066_F7028PGW

With (A) as the document from the CN

1. Verified that the size of the object differs from that in the system metadata, as does the checksum.
2. Verified that the object is not retrievable from the member node (node offline)
3. Google search on the PID shows the ONEMercury interface, no Google search results to the clearing house.
4. The DOI listed in the metadata is "doi:10.5066/F7028PGW" The DOI resolves to a zip file that contains an ArcGIS layer and a PDF document, no metadata.
5. The URL in the metadata resolves to an fgdc file (B).
6. Searching for the DOI in the USGS search box returns one result, which points us to: www1.usgs.gov/vip/kaho/metakahospatial.xml downloaded as (C)
7. diff reports no difference between (B) and (C)
8. diff reports minor differences between (A) and (B) (< is copy (A) from CN):

1,2c1,2
< <?xml version="1.0"?>

< http://www1.usgs.gov/metadata/mdata/vip/metakahospatial.xml

> <?xml version="1.0" encoding="ISO-8859-1"?>
>
> 30c30

< Cogan, D. K. Schulz., D. Benitez, G. Kudray, and A. Ainsworth 2011. Vegetation inventory project: Kaloko-Honokohau National Historical Park NPS/KAHO/NRR 2011/462. National Park Service, Fort Collins, Colorado.

> Cogan, D. K. Schulz., D. Benitez, G. Kudray, and A. Ainsworth 2011. Vegetation inventory project: Kaloko-Honokohau National Historical Park NPS/KAHO/NRR2011/462. National Park Service, Fort Collins, Colorado.

Conclusions:

- the original content is not available in exactly the same form as published to the CN.
- the currently available content differs from that on the CN in a primarily cosmetic manner.
- No currently available copies report the same size or checksum as recorded in the system metadata.
- Adding the element to (B) did not reconcile reconcile the difference of checksum and size from (A)

Hence, there is by definition, no valid copy of the original data from DataONE's perspective. From a pragmatic viewpoint, the content remains available at the locations referenced within the metadata document (A), and so is still practically useful.

From a user perspective, the content remains valid. From a user perspective, the checksum and size should be updated to reflect that the copy held by the CN is the only valid copy. Since this is a version 1.0 object, such a change is not possible without violating self imposed integrity constraints.

One possible solution may be to "upgrade" the system metadata to version 2.0, make the current PID the SID, and create a new system metadata document to indicate the current state of the object, and reference the original system metadata entry as obsoleted.

#5 - 2016-01-26 19:29 - Skye Roseboom

- Target version set to CCI-2.2.0

#6 - 2016-01-26 19:29 - Skye Roseboom

- Target version deleted (CCI-2.2.0)