Gmail by Google

Mark Servilla <mark.servilla@gmail.com>

Example of broken object in Metacat

6 messages

Roger Dahl <dahl@unm.edu> Mon, Aug 11, 2014 at 9:22 AM
To: Mark Servilla <mservilla@lternet.edu>

sys_meta.size = 10979
sys_meta_checksum = f8fb6c6eb0558cc5408aeff5bf71eb72
sys_meta_checksum_algo = MD5

real_len = 13046
real_md5 = da1fa6a56f8e602656dd2ae13866253e

Attached, the raw sysmeta and object bytes.

2 attachments

doi_10_6073_AA_cce_155_1
14K

doi_10_6073_AA_cce_155_1.sysmeta
1K

Roger Dahl <dahl@unm.edu> Mon, Aug 11, 2014 at 11:27 AM
To: Christopher Jones <cjones@nceas.ucsb.edu>, Jing <tao@nceas.ucsb.edu>, Robert Waltz <rwaltz@utk.edu>
Cc: Mark Servilla <mservilla@lternet.edu>

Total values:

Objects checked: 60743
Objects with errors: 36287
Checksum mismatches: 36263
Size mismatches: 36256
System Metadata errors: 1 (after fudging)
Science Object errors: 23

The "after fudging" is because there's another small issue with the science data that cause them not to validate against the schema. It's that the sysmeta contains empty accessPolicy and blockedMemberNode elements. Those elements must be populated if they're present. So I read the sysmeta as a string and remove the empty elements before parsing it.

See below for info about the object that is attached. I'm generating the complete csv now.

Roger
[Quoted text hidden]

2 attachments

doi_10_6073_AA_cce_155_1
14K

doi_10_6073_AA_cce_155_1.sysmeta
1K

Mark Servilla <servilla@lternet.edu> Tue, Aug 12, 2014 at 12:36 PM
Reply-To: servilla@lternet.edu
To: Roger Dahl <dahl@unm.edu>, Christopher Jones <cjones@nceas.ucsb.edu>, Jing <tao@nceas.ucsb.edu>, "Waltz, Robert Patrick" <rwaltz@utk.edu>, Ben Leinfelder <leinfelder@nceas.ucsb.edu>

Hi All,

The following is a brief analysis of the discrepancy we are seeing between science metadata objects from Tropical (LTER's Metacat) and the CN when looking at one specific example for the science metadata object with identifier "doi:10.6073/AA/cce.155.1":

System metadata size and checksum are identical between Tropical (Metacat) and the CN. Note difference when system metadata was updated between Tropical and CN; CN has newer version of system metadata.

From tropical -

```
curl -i -k -X GET https://tropical.lternet.edu/knb/d1/mn/v1/meta/doi%3A10.
6073%2FAA%2Fcce.155.1 -E ./urn_node_LTER.pem

<?xml version="1.0" encoding="UTF-8"?>
<d1:systemMetadata xmlns:d1="http://ns.dataone.org/service/types/v1">
    <serialVersion>1</serialVersion>
    <identifier>doi:10.6073/AA/cce.155.1</identifier>
    <formatId>eml://ecoinformatics.org/eml-2.1.0</formatId>
    <size>10979</size>
    <checksum algorithm="MD5">f8fb6c6eb0558cc5408aeff5bf71eb72</checksum>
    <submitter>uid=CCE,o=LTER,dc=ecoinformatics,dc=org</submitter>
    <rightsHolder>uid=CCE,o=LTER,dc=ecoinformatics,dc=org</rightsHolder>
    <accessPolicy/>
    <replicationPolicy replicationAllowed="false"/>
    <archived>true</archived>
    <dateUploaded>2010-10-21T23:00:00.000+00:00</dateUploaded>
    <dateSysMetadataModified>2012-06-26T00:36:57.140+00:00</dateSysMetadataModified>
    <originMemberNode>urn:node:LTER</originMemberNode>
    <authoritativeMemberNode>urn:node:LTER</authoritativeMemberNode>
</d1:systemMetadata>
```

From CN -

```
curl -i -k -X GET https://cn.dataone.org/cn/v1/meta/doi%3A10.6073%2FAA%2Fcce.155.1 -
E ./urn_node_LTER.pem

<?xml version="1.0" encoding="UTF-8"?>
<d1:systemMetadata xmlns:d1="http://ns.dataone.org/service/types/v1">
    <serialVersion>1</serialVersion>
    <identifier>doi:10.6073/AA/cce.155.1</identifier>
    <formatId>eml://ecoinformatics.org/eml-2.1.0</formatId>
    <size>10979</size>
    <checksum algorithm="MD5">f8fb6c6eb0558cc5408aeff5bf71eb72</checksum>
    <submitter>uid=CCE,o=LTER,dc=ecoinformatics,dc=org</submitter>
    <rightsHolder>uid=CCE,o=LTER,dc=ecoinformatics,dc=org</rightsHolder>
    <accessPolicy/>
    <replicationPolicy replicationAllowed="false"/>
    <archived>true</archived>
    <dateUploaded>2010-10-21T23:00:00.000+00:00</dateUploaded>
    <dateSysMetadataModified>2013-06-14T16:12:50.809+00:00</dateSysMetadataModified>
```

```
    <originMemberNode>urn:node:LTER</originMemberNode>
    <authoritativeMemberNode>urn:node:LTER</authoritativeMemberNode>
    <replica>
        <replicaMemberNode>urn:node:CN</replicaMemberNode>
        <replicationStatus>completed</replicationStatus>
        <replicaVerified>2012-06-26T00:00:00.000+00:00</replicaVerified>
    </replica>
    <replica>
        <replicaMemberNode>urn:node:LTER</replicaMemberNode>
        <replicationStatus>completed</replicationStatus>
        <replicaVerified>2012-06-26T00:00:00.000+00:00</replicaVerified>
    </replica>
</d1:systemMetadata>
```

There are, however, big differences between the two versions of science metadata (1) retrieved from Tropical, (2) that exists on Tropical's filesystem, and (3) retrieved from the CN. First, the science metadata retrieved from Tropical and that exists on Tropical's filesystem (as part of the Metacat system) is identical in size, and it jibes with what Roger's Python script shows (see below; the first line is what exists on Tropical's filesystem and the second line is retrieved from Tropical using the MN get object call). The third line shows the science metadata object retrieved from the CN using the get object call; the size of this object is the same as declared in both sets of system metadata. Note that the corresponding curl commands are below.

```
-rw-r--r-- 1 tcat tcat 13046 2010-10-21 20:00 cce.155.1

-rw-r--r--  1 servilla  staff  13046 Aug 12 12:05 cce.155.1.tropical.xml

-rw-r--r--  1 servilla  staff  10979 Aug 12 12:06 cce.155.1.cn.xml

curl -k -X GET https://tropical.lternet.edu/knb/d1/mn/v1/object/doi%3A10.
6073%2FAA%2Fcce.155.1 -E ./urn_node_LTER.pem > cce.155.1.tropical.xml

url -k -X GET https://cn.dataone.org/cn/v1/object/doi%3A10.6073%2FAA%2Fcce.155.1 -E
./urn_node_LTER.pem > cce.155.1.cn.xml
```

The significant differences between the two science metadata objects is that the version (both) from Tropical contains newline and whitespace characters (pretty XML formatting), where as the version from the CN does not, and the version from the CN contains a XML attribute "scope="document" in a number of elements, where as the version from Tropical does not. As such, the difference in size must be attributed to the newline and whitespace found in the version from Tropical and the attribute found in the version from the CN, with the system metadata being based on the version replicated to the CN.

Now, the question I have is why is there a difference between the two science metadata objects between Tropical and the CN?

Attachments are the corresponding science metadata objects.

Mark Servilla, Ph.D.

LTER Network Office
Department of Biology
MSC 03 2020
1 University of New Mexico
Albuquerque, NM 87131-0001

servilla@LTERnet.edu
(505) 750-3226
[Quoted text hidden]

2 attachments

cce.155.1.cn.xml
11K

cce.155.1.tropical.xml
13K

Benjamin Leinfelder <leinfelder@nceas.ucsb.edu> Tue, Aug 12, 2014 at 5:18 PM
To: servilla@lternet.edu
Cc: Roger Dahl <dahl@unm.edu>, Christopher Jones <cjones@nceas.ucsb.edu>, Jing <tao@nceas.ucsb.edu>, "Waltz, Robert Patrick" <rwaltz@utk.edu>

Hi Mark,
This sounds like something we were seeing in Metacat, pre-1.9.x release, and prompted us to store the original XML metadata content on disk rather than relying on a parsed/decomposed version in the Metacat DB.
I'm not sure why this would be flaring up now on the CNs, and may not even be the actual cause of the discrepancy. My only guess would be that there was an issue when trying to write the original to disk initially and we had to fall back to the DB-stored version of the file at some point.

Options for correcting this (assuming we have an expectation it won't happen again for the same content) might be to delete the affected objects entirely and force synchronization to run again. Or ignore these and move on?
-ben
[Quoted text hidden]
> <cce.155.1.cn.xml><cce.155.1.tropical.xml>

Mark Servilla <servilla@lternet.edu> Tue, Aug 12, 2014 at 9:22 PM
Reply-To: servilla@lternet.edu
To: Benjamin Leinfelder <leinfelder@nceas.ucsb.edu>
Cc: Roger Dahl <dahl@unm.edu>, Christopher Jones <cjones@nceas.ucsb.edu>, Jing <tao@nceas.ucsb.edu>, "Waltz, Robert Patrick" <rwaltz@utk.edu>

Thanks for the feedback, Ben. Perhaps we can discuss this issue a bit more at the end of stand-up tomorrow.

Sincerely,
Mark

Mark Servilla, Ph.D.

LTER Network Office
Department of Biology
MSC 03 2020
1 University of New Mexico
Albuquerque, NM 87131-0001

servilla@LTERnet.edu
(505) 750-3226

[Quoted text hidden]

Mark Servilla <servilla@lternet.edu> Wed, Aug 13, 2014 at 2:44 PM
Reply-To: servilla@lternet.edu
To: Benjamin Leinfelder <leinfelder@nceas.ucsb.edu>

Cc: Roger Dahl <dahl@unm.edu>, Christopher Jones <cjones@nceas.ucsb.edu>, Jing <tao@nceas.ucsb.edu>, "Waltz, Robert Patrick" <rwaltz@utk.edu>

Gentlemen,

The situation is a bit stranger than first suspected. Roger provided an example science metadata identifier of an object that did not trigger the size/checksum error. Unfortunately, this object poses new issues:

1. The system metadata size value and the object size (both downloaded using the MN API and on filesystem) on the Metacat MN jibe.

```
<?xml version="1.0" encoding="UTF-8"?>
<d1:systemMetadata xmlns:d1="http://ns.dataone.org/service/types/v1">
    <serialVersion>1</serialVersion>
    <identifier>doi:10.6073/AA/cdonahue.18.2</identifier>
    <formatId>eml://ecoinformatics.org/eml-2.0.1</formatId>
    <size>15060</size>
    <checksum algorithm="MD5">54438724c4d6cc04a5424df1d4de1ffa</checksum>
    <submitter>uid=sbc,o=LTER,dc=ecoinformatics,dc=org</submitter>
    <rightsHolder>uid=cdonahue,o=LTER,dc=ecoinformatics,dc=org</rightsHolder>
    <accessPolicy/>
    <replicationPolicy replicationAllowed="false"/>
    <obsoletes>cdonahue.18.1</obsoletes>
    <archived>true</archived>
    <dateUploaded>2009-06-22T23:00:00.000+00:00</dateUploaded>
    <dateSysMetadataModified>2012-07-01T21:40:39.372+00:00</dateSysMetadataModified>
    <originMemberNode>urn:node:LTER</originMemberNode>
    <authoritativeMemberNode>urn:node:LTER</authoritativeMemberNode>
</d1:systemMetadata>

-rw-r--r--  1 servilla  staff  15060 Aug 13 14:40 cdonahue.18.2.tropical.xml
```

2. The system metadata size value and object size on the CN jibe

```
<?xml version="1.0" encoding="UTF-8"?>
<d1:systemMetadata xmlns:d1="http://ns.dataone.org/service/types/v1">
    <serialVersion>1</serialVersion>
    <identifier>doi:10.6073/AA/cdonahue.18.2</identifier>
    <formatId>eml://ecoinformatics.org/eml-2.0.1</formatId>
    <size>16919</size>
    <checksum algorithm="MD5">f0b2eb4810fb49ca1c91adf1899f3560</checksum>
    <submitter>uid=sbc,o=LTER,dc=ecoinformatics,dc=org</submitter>
    <rightsHolder>uid=cdonahue,o=LTER,dc=ecoinformatics,dc=org</rightsHolder>
    <accessPolicy/>
    <replicationPolicy replicationAllowed="false"/>
    <obsoletes>cdonahue.18.1</obsoletes>
    <archived>false</archived>
    <dateUploaded>2009-06-22T23:00:00.000+00:00</dateUploaded>
    <dateSysMetadataModified>2012-06-26T07:36:38.210+00:00</dateSysMetadataModified>
    <originMemberNode>urn:node:LTER</originMemberNode>
    <authoritativeMemberNode>urn:node:LTER</authoritativeMemberNode>
    <replica>
        <replicaMemberNode>urn:node:CN</replicaMemberNode>
        <replicationStatus>completed</replicationStatus>
        <replicaVerified>2012-06-26T00:00:00.000+00:00</replicaVerified>
    </replica>
    <replica>
        <replicaMemberNode>urn:node:LTER</replicaMemberNode>
```

```
        <replicationStatus>completed</replicationStatus>
        <replicaVerified>2012-06-26T00:00:00.000+00:00</replicaVerified>
    </replica>
</d1:systemMetadata>

-rw-r--r--@ 1 servilla  staff  16919 Aug 13 13:53 cdonahue.18.2.cn.xml
```

3. The system metadata size value and the objects differ between the Metacat MN and the CN. It appears that the CN object is, again, without any whitespace, but it is noticeably larger - probably due again to the introduction of the scope="document" attribute, which occurs 39 times, in the CN version of the object.

Sincerely,
Mark

Mark Servilla, Ph.D.

LTER Network Office
Department of Biology
MSC 03 2020
1 University of New Mexico
Albuquerque, NM 87131-0001

servilla@LTERnet.edu
(505) 750-3226