Metacat - Bug #3296

Replication: Many EML documents fail to replicate

05/13/2008 02:37 PM - Duane Costa

Status:	Resolved	**Start date:**	05/13/2008
Priority:	Immediate	**Due date:**	
Assignee:	Jing Tao	**% Done:**	0%
Category:	metacat	**Estimated time:**	0.00 hour
Target version:	2.0.0	**Spent time:**	0.00 hour
Bugzilla-Id:	3296		

Description

Over 1000 EML documents fail to replicate from LTER Metacat to KNB Metacat during every timed replication, currently scheduled once per week. Jing and I did some initial investigation of this issue last year. Jing suspects that the problem might be related to normalization (and/or lack of normalization) of character entities such as '<'. (See below.)

A secondary bug seems to be that each of the failed documents is sent twice, not once. (Also described below.) Thus, for each timed replication, LTER attempts to send over 2000 documents to KNB, KNB fails to write them into its Metacat, and the process is repeated again the following week. Because of the large number of documents sent during each timed replication, the process requires about eleven hours to complete.

We were previously scheduling timed replications once every 48 hours, but as an interim measure we reduced this to once every 7 days to reduce the load on both Metacat servers.

On Tue, 17 Jul 2007, Jing Tao wrote:

Here is something on the top of my head (if anybody know it is wrong, please correct me):

When you send xml document containing something like "<" in text part, xerces will automatically transform it to "<" during our metacat inserting process. So "<" will be stored in our db. When we try to read the xml file, metacat should transform the stored "<" to "<". Otherwise, the xml file will be not well-formed. The details of the transform (normalize) is in MetacatUtil.normalize().

> The documents that are failing to be replicated are being accessed twice >every forty-eight hours. This is strange, because we expected them to be >accessed only once every forty-eight hours. Here's an example:
>
> > grep knb-lter-and.3190.4 metacatreplication.log
>
> 07-07-09 01:06:01 :: document knb-lter-and.3190.4 sent
> 07-07-09 06:29:32 :: document knb-lter-and.3190.4 sent
> 07-07-11 01:06:39 :: document knb-lter-and.3190.4 sent
> 07-07-11 06:25:31 :: document knb-lter-and.3190.4 sent
> 07-07-13 01:07:34 :: document knb-lter-and.3190.4 sent
> 07-07-13 07:03:49 :: document knb-lter-and.3190.4 sent
> 07-07-15 01:09:06 :: document knb-lter-and.3190.4 sent
> 07-07-15 07:34:07 :: document knb-lter-and.3190.4 sent
>
> You can see how LNO attempts to send the same document twice, not once, about six hours apart every two days. Why is this happening, shouldn't it only be attempted once every forty-eight hours?

I took a look at the replication log files and found the time interval is 48 hours, but somehow the replication request are called twice in one replication! It is a bug.

> Jing Tao wrote:
>
> > Hi, Duane:
> >
> > Thank you for the info. I took a look at the error log in knb site. Yeah, knb rejected lots of documents from LNO. From the error message, it seems those documents have some problem in format. Since LNO can accecept them, I don't think there is any problem in oringin docs. The most possibile problme, which I am guessing, is the normalization. Xerces parsor automatically normailize some stuff. However, metacat failed to denormalize the doc when we try to read it. How do you

think? I will take a look soon. The attached files are replication error and log in knb.

Thanks,

Jing

> Hi Jing,
>
> I've attached the LNO replication log files from the past few days. It looks like LNO is replicating many files to KNB with every timed replication, but the KNB server is not accepting them. Also, replication attempt seems to take long, about four or five minutes per document. I think maybe the timed replications were beginning to take so long that they even started to overlap, i.e. a new timed replication would begin even though the previous one from forty-eight hours ago hadn't completed yet. When you get a chance could you please take a look at the KNB logs to match this up and find out why the documents are not being accepted on the KNB side?
> Thanks,
> Duane
>
> On 7/11/2007, Duane Costa wrote:
>
> > Hi Jing,
> >
> > I was just checking the replication logs on the LNO metacat, and it looks like there is an unusually large amount of replication happening with the timed replications (once every 48 hours), with many documents being sent from LNO to KNB. Do you know of any changes (on the KNB side) that might be causing this, and could you check the replication logs at the KNB metacat to see if you agree that it looks unusual?
> >
> > Thanks,
> > Duane

History

#1 - 05/13/2008 03:15 PM - Jing Tao

I almost forgot this bug! Thanks for input it. Can you join to irc on tomorrow?

#2 - 05/14/2008 11:01 AM - Jing Tao

it seems we need to correct local documents in lter. it has no relation to the release. So i postpone it to 1.9

#3 - 05/14/2008 01:29 PM - Duane Costa

Jing and I investigated further. Here's what we've learned so far:

(1) Many of the EML documents that fail to be replicated are old revisions. In many cases the newer revisions have successfully been replicated from LTER to KNB. This could indicate that something that was broken in Metacat when the older EML revision was inserted into LTER has since been fixed in a newer version of Metacat when the newer revision of the EML document was inserted. Or it could mean that something in the newer revision of the EML document itself changed to overcome the problem (less likely, we think).

I don't have an exact percentage of which EML documents that fail to replicate are old revisions versus current revisions. It would take a lot of time to do this analysis. A sampling of EML documents that cannot be replicated seems to indicate that many are old revisions. This is good, since it's less important for old revisions to be replicated (ideally, of course, we would like everything to be replicated for completeness). On the other hand, it's bad that these replication failures continue to be repeated with every timed replication, using up resources on both the LTER and KNB metacat servers.

(2) We compared an EML document in LTER that fails to be replicated to its original source document (in this case at the CDR LTER site). We found differences such as:

|Exp|Plot|Subplot|taxon|Species|Seedlings <= Labels

|Exp|Plot|Subplot|taxon|Species|Seedlings <= Labels

which involve XML character entities. In this example, the original document contained the character entity '<', while the Metacat document contained the literal '<' character, causing it to fail XML validation, which explains why it cannot be replicated from LTER to KNB.

Jing thinks that the differences were introduced by a known bug which has since been resolved in Metacat 1.8.1:

http://bugzilla.ecoinformatics.org/show_bug.cgi?id=2517

(3) Possible resolutions. Jing suggests manual changes to the LTER Metacat database could fix these problems. For example, find all literal instances of '<' in the nodedata field of xml_nodes and xml_nodes_revisions and change them to the character entity '<'. (Likewise for '>' and perhaps other characters, but need to be careful about '&'.) We both agree that any changes to the content of the nodedata field would have to be done with extreme caution because:
(a) changing nodedata is a big deal, it means you're changing the actual content of the document;
(b) you could easily end up making things worse instead of better!

Duane
(

#4 - 05/16/2008 03:35 PM - Duane Costa

Some additional analysis of the replication errors:

A total of 1016 EML documents fail to replicate from LNO to KNB.

668 of the documents that fail to replicate are old revisions (i.e. in xml_revisions table).

348 of the documents that fail to replicate are current revisions (i.e. in xml_documents table).

#5 - 05/21/2008 09:04 AM - Duane Costa

Of the 1016 EML documents that fail to replicate from LNO to KNB, we identified 611 that are directly related to literal characters in inline data files that resulted in invalid XML. There were two cases:

1. literal '&' character in inline data, e.g. :

```
'Mosses & Lichens'
```

2. literal '<' character in inline data, e.g. :

```
'|Exp|Plot|Subplot|taxon|Species|Seedlings <= Labels
```

Of these 611 instances, 254 were in current EML revisions and 357 were in old EML revisions. All 611 documents originated from the Cedar Creek (CDR) LTER and all document identifiers began with the string pattern 'knb-lter-cdr'. It appears that the original CDR documents contained the appropriate XML character entities but these were converted to literal '&' and '<' characters during insertion into Metacat (possibly related to bugs #2517 and #2797).

We have repaired all 611 of these documents using the following procedure:

1. For safety, we created a backup of the 'data/knb/inlinedata' directory.

2. In the original 'data/knb/inlinedata' directory, we ran the following Perl commands:

```
(a) perl -pi -e 's/ \& / \&amp; /g' *
```

```
(b) perl -pi -e 's/</&lt;/g' *
```

Note that the first Perl command included a leading and trailing space character in the string pattern. This was a precautionary measure to avoid replacing the leading ampersand character in existing character entities (which could result in something incorrect like '&amplt;'); we were able to use this pattern because all instances of literal ampersands that needed replacement were of the form ' & '.

This repair procedure has two benefits:

1. There were 254 CDR documents that appeared in Metacat search results in the LNO Metacat, but they could not be read because they generated invalid XML. All 254 of these documents should now be able to be read.

2. All 611 CDR documents (254 current, 357 old) that previously could not be replicated to knb.ecoinformatics.org should now replicate successfully. The next timed replication from LNO to KNB is expected to run this evening (5/21/2008) so we should be able to verify this tomorrow morning.

#6 - 06/05/2008 01:46 PM - Duane Costa

We have reduced the number of EML documents that fail to replicate from the original 1016 down to 94. This was accomplished by:

1. Repair of inline data for 611 CDR documents as described in the previous comment. Note, however, that although these documents were successfully replicated to KNB, they also needed to be repaired at KNB because the same Metacat bug that corrupted them when they were inserted at LTER has also corrupted them at KNB.

2. Purge of 311 corrupted documents from the LTER metacat. All 311 of these documents were old revisions, not current revisions. (LTER Information Managers at the affected sites were given prior notice of the plan to purge corrupted old revisions and there were no objections.)

Some ongoing concerns include:

1. There are still 94 documents that fail to replicate. These are all current revisions. In cases where we know how a document can be repaired, we will work with the LTER Information Manager to correct the problem, as time allows.

2. There are 188 replication errors because each of the 94 documents fails to replicate twice. This replication bug (#3304) has already been fixed by Jing, but the fix has not yet been deployed to the KNB metacat.

3. Timed replication requires almost eight hours to complete. (Of course, this should be cut in half after the bug fix for #3304 is deployed.) We'd like to understand better why it takes so long to replicate 188 documents.

4. The 611 CDR documents replicated to KNB need to be repaired by fixing their inline data.

#7 - 11/09/2011 09:41 AM - Duane Costa

Follow-up to the previous comment of 2008-06-05:

1. There are still 94 documents that fail to replicate. These are all current
revisions. In cases where we know how a document can be repaired, we will work
with the LTER Information Manager to correct the problem, as time allows."

-- This is an ongoing LTER issue so it doesn't need to be tracked in this bug.

2. There are 188 replication errors because each of the 94 documents fails to
replicate twice. This replication bug (#3304) has already been fixed by Jing,
but the fix has not yet been deployed to the KNB metacat.

-- Replication log output indicates that replication is now running once instead of twice, verifying that Jing's bug fix is now deployed.

3. Timed replication requires almost eight hours to complete. (Of course, this
should be cut in half after the bug fix for #3304 is deployed.) We'd like to
understand better why it takes so long to replicate 188 documents.

-- Timed replication now completes in less than ninety minutes.

4. The 611 CDR documents replicated to KNB need to be repaired by fixing their
inline data.

-- Jing will open a separate bug to track this issue.

This bug can now be closed out.

#8 - 03/27/2013 02:23 PM - Redmine Admin

Original Bugzilla ID was 3296