Metacat - Bug #3241

Update older invalidated eml201 document in Metacat

04/24/2008 02:59 PM - Jing Tao

Status:	Resolved	**Start date:**	04/24/2008
Priority:	Immediate	**Due date:**	
Assignee:	Jing Tao	**% Done:**	0%
Category:	metacat	**Estimated time:**	0.00 hour
Target version:	1.8.1	**Spent time:**	0.00 hour
Bugzilla-Id:	3241		

Description

Previous Metacat release shipped wrong eml201 schema. See bug 3209. If we move metacat to point to correct schema, some existed eml documet may be invalidate. We need a batch file to correct this issue.

History

#1 - 05/02/2008 03:59 PM - Jing Tao

To my understanding, Attribute "system="document"" in element "references" is the only extra stuff which need to be removed.

Here is my sql commands:
1. delete from xml_index where doctype ='eml://ecoinformatics.org/eml-2.0.1' AND nodeid in (select nodeid from xml_index where path ='references/@system');

2. delete from xml_nodes where nodetype='ATTRIBUTE' and nodename='system' and nodedata='document' and parentnodeid in (select nodeid from xml_nodes where nodetype='ELEMENT' and nodename='references') and docid in (select docid from xml_documents where doctype ='eml://ecoinformatics.org/eml-2.0.1');

it seems working well.

However there is an issue (maybe not):
In xml_nodes table
before deleting, we will have something like:

```
nodeindex  nodetye nodename nodedata
1          attribute system  document
2          text              1234
```

after deleting, the first row will be gone. So we have nodeindex 2 in table, but no nodeindex 1 in table.

So far, i didn't see any issue in read and query. But do you think it will cause any problem?

#2 - 05/05/2008 08:46 AM - Matt Jones

This will probably be ok, unless you delete a node that has a child node in the table. Does the 'system' ATTRIBUTE node have a TEXT child node? You should also check DocumnetImpl.read* to be sure that the node selection and document reconstruction algorithm allows the nodes to be non-sequential in their IDs. Right now I think it assumes they are in order but need not be sequential.

#3 - 05/08/2008 04:40 PM - Jing Tao

In oder to make sure 'system' ATTRIBUTE node doesn't have any child node, I ran this query in knb server:

```
select count(*) from xml_nodes where parentnodeid in (select nodeid from xml_nodes where nodetype='ATTRIBUTE' and nodename='system' and nodedata='document'and parentnodeid in (select nodeid from xml_nodes where nodetype='ELEMENT' and nodename='references') and docid in (select docid from xml_documents where doctype ='eml://ecoinformatics.org/eml-2.0.1'));
count
-------
0
(1 row)
```

It seem there is no child node of this attribute.

Moreover, there is a foreign key in xml_nodes table:
"xml_nodes_parent_fk" FOREIGN KEY (parentnodeid) REFERENCES xml_nodes(nodeid)
If 'system' attribute has a child node, i couldn't delete the attribute successfully in my test machine, which has an early version (2007)of knb db.

Matt, i was talking about nodeidex in my previous comment. However you pointed out a more important issue - nodeid.

Here is some result after I went through DocumentImpl.toXml method:

1) Read algorithm doesn't use nodeindex any more. I commented out the method getNodeIndex in NodeRecord class, the compile errors showed that only couple debug statements use this method.

2)The nodeid should be ordered but not necessary be sequential. Here is only place where DocumentImpl.toXml method uses getParenNodeId and getNodeId methods.

```
if (currentNode.getParentNodeId() != currentElement.getNodeId()) {
while (currentNode.getParentNodeId() != currentElement.getNodeId()) {
```

Those methods are used to determine the relationship between parents and children. We don't assume nodeid should be sequential.

#4 - 05/12/2008 02:33 PM - Jing Tao

Developed a new class named EML201DocumentCorrector class in metacat.

This class will run those three sql commands in order:
1. Deleting in xml_index table:
delete from xml_index where doctype ='eml://ecoinformatics.org/eml-2.0.1' AND nodeid in (select nodeid from xml_index where path ='references/@system');

2. Deleting in xml_nodes table:
delete from xml_nodes where nodetype='ATTRIBUTE' and nodename='system' and nodedata='document' and parentnodeid in (select nodeid from xml_nodes where nodetype='ELEMENT' and nodename='references') and docid in (select docid from xml_documents where doctype ='eml://ecoinformatics.org/eml-2.0.1');

3. Deleting in xml_nodes_revisions table:
delete from xml_nodes_revisions where nodetype='ATTRIBUTE' and nodename='system' and nodedata='document' and parentnodeid in (select nodeid from xml_nodes where nodetype='ELEMENT' and nodename='references') and docid in (select docid from xml_documents where doctype ='eml://ecoinformatics.org/eml-2.0.1');

#5 - 05/12/2008 04:16 PM - Jing Tao

Ran the above class file in dev.nceas. It took 2 minutes to finish the task successfully.

#6 - 05/14/2008 02:57 PM - Jing Tao

The new class will put into initialize method in MetacatServlet class. A new property named eml201_document_corrected with default value "false" was added into metacat.properties. If this value is false, metacat will run the Corrector class and set eml201_document_corrected to "true" if the process succeed. So when next time you start metacat, it would not rerun this class again.

#7 - 05/14/2008 02:59 PM - Jing Tao

Margaret confirmed that there only one difference between the two versions of eml-resource.xsd;
Hi Jing -
I did the most recent work on eml-resource.xsd. Yes, the only difference between 1.78 (in RELEASE_EML_2_0_1) and 1.79 (in several later UPDATES) is the presence of
the "system" attribute on the "references" element that includes a nonsense default. This is explained in a comment to bug 1662, which is the bug that r1.79's changes
addressed.

I put those changes into an existing branch and reverted back to 1.78 for r1.80. Subsequent check ins to the resource.xsd were only to update the document declaration
for 2.1.0-release-candidates, so the current revision is now at 1.83.

margaret

Jing Tao wrote:

> Hi,devs:
>
> I just went through the **.xsd file in eml and got the same conclusion as before that tag RELEASE_EML_2_0_1 and tag RELEASE_EML_2_0_1_UPDATE_** point same version of
> .xsd files except eml-resource.xsd.
>
> RELEASE_EML_2_0_1 and RELEASE_EML_2_0_1_UPDATE_1 point version 1.78 of eml-resource.xsd file, but from RELEASE_EML_2_0_1_UPDATE_2 to RELEASE_EML_2_0_1_UPDATE_6
> point version 1.79.
>
> Currently we know there is a difference in "system" attribute in "references" element. Here is the comment in 1.79 check-in:modified referenceGroup to add support
> for system attribute.
>
> Is this the only difference between the two versions?
>
> I tried to use "cvs diff" to help me out, but I got lots of difference lines. Before I go through every line, i think it will be good to ask devs to give me

some
hint. Do you know any difference there?

Thank you!

Jing

#8 - 06/03/2008 09:26 AM - Callie Bowdish

There is the possibility of older documents with illegal eml trying to be saved to Metacat. If these older versions have illegal eml elements in them there will be an error. The end user working with Morpho, trying to save an illegal eml data package, will not see the error. They will just notice that the data package did not save to the network.

The 1.8.1 release of Metacat fixes an EML error in the element 'references' that documents saved up to Metacat in an earlier release may contain. If a document, that has been downloaded from an earlier version of Metacat, contains the element that looks like this <references system="document"> there will be an error when trying to save it to the new release of Metacat.

The error is: cvc-type.3.1.1: [attributes] of element 'references' must be empty, excepting those whose [namespace name] is identical to http://www.w3.org/2001/XMLSchema-instance and whose [local name] is one of type, nil, schemaLocation or noNamespaceSchemaLocation.

Documents on Metacat 1.8.1 have had the illegal attribute removed by a script that runs upon instillation. If you want to save a data set with this error you will need to remove the system="document" part of the references tag.

#9 - 03/27/2013 02:22 PM - Redmine Admin

Original Bugzilla ID was 3241