Metacat - Bug #5696

pathQuery returns eml docs which have no public access granted

08/24/2012 12:06 PM - gastil gastil

Status:	Resolved	**Start date:**	08/24/2012
Priority:	Normal	**Due date:**	
Assignee:	ben leinfelder	**% Done:**	0%
Category:	metacat	**Estimated time:**	0.00 hour
Target version:	2.0.5	**Spent time:**	0.00 hour
Bugzilla-Id:	5696		

Description

As far as I remember, non-public eml docs did not used to be returned in pathQuery result sets in earlier versions of metacat.
This is with
http://metacat.lternet.edu/knb/metacat?action=getversion
<version>2.0.3</version>

A pathQuery returns an eml doc which does not have public read access.
Example: knb-lter-sev.389.3

with
<access authSystem="knb" order="denyFirst" scope="document">
<allow>
<principal>uid=SEV, o=lter, dc=ecoinformatics, dc=org</principal>
<permission>all</permission>
</allow>
</access>

A pathQuery returned this in its result set:
<document>
<docid>knb-lter-sev.389.3</docid>
<docname>eml</docname>
<doctype>eml://ecoinformatics.org/eml-2.0.1</doctype>
<createdate>2005-07-29</createdate>
<updatedate>2012-08-22</updatedate>
<param name="@packageId">sev.00389.1</param>
<param name="dataset/title">Lightning Strike Data for New Mexico, 1989</param>
</document>

This may be related in part to bug #5553 (not sure).
The denyFirst may be part of the problem. The older revisions also had denyFirst.

History

#1 - 09/03/2012 11:20 AM - ben leinfelder

I've recreated the scenario on my local Metacat installation -- it appears the permissions for the older revision are still being applied to the newer revision in the search results. I suspect this is related to how the index is purged, or not, as the case seems to indicate.

#2 - 09/04/2012 08:43 AM - ben leinfelder

Metacat v2.0.4 will include a fix for this issue.

#3 - 09/28/2012 08:11 AM - ben leinfelder

Seems this query can be quite expensive when the DB has a large number of documents. Re-working to remove the max(rev) condition - hoping that it does not require a massive overhaul of the QuerySpecification->SQL code.

#4 - 09/28/2012 09:20 AM - ben leinfelder

With the join to the xml_documents table, the response is better - but not that great (3 minutes for "tree" keyword serarch:

MetacatHandler.handleSQuery - squery:
<pathquery version="1.2">
<querytitle>Advanced Search</querytitle>

<returnfield>keyword</returnfield>
<returndoctype>eml://ecoinformatics.org/eml-2.1.0</returndoctype>
<returndoctype>eml://ecoinformatics.org/eml-2.0.1</returndoctype>
<returndoctype>eml://ecoinformatics.org/eml-2.0.0</returndoctype>
<querygroup operator="UNION">
<queryterm searchmode="contains" casesensitive="false">
<value>tree</value>
<pathexpr>keyword</pathexpr>
</queryterm>
</querygroup>
</pathquery>
ran in 179648 ms [edu.ucsb.nceas.metacat.MetacatHandler]

#5 - 09/28/2012 09:31 AM - ben leinfelder

The expensive part seems to be the subqueries of all public-read docs and all public-read-deny docs:

SELECT docid,docname,doctype,date_created, date_updated, rev FROM xml_documents WHERE docid IN ((SELECT DISTINCT docid FROM xml_path_index WHERE ((UPPER LIKE TREE AND path LIKE keyword)))) AND (docid IN (SELECT id.docid from xml_access xa, identifier id, xml_documents xmld WHERE id.guid = xa.guid AND id.docid = xmld.docid AND id.rev = xmld.rev AND ((lower(principal_name) = 'public') AND perm_type = 'allow' AND permission > 3)) AND docid NOT IN (SELECT id.docid from xml_access xa, identifier id, xml_documents xmld WHERE id.guid = xa.guid AND id.docid = xmld.docid AND id.rev = xmld.rev AND ((lower(principal_name) = 'public') AND perm_type = 'deny' AND perm_order ='allowFirst' AND permission > 3)))

#6 - 10/18/2012 04:40 PM - ben leinfelder

I've reworked how access was being checked -- now we have a simpler clause there and the current revision handling is done "higher up" in the query -- this saves us a lot of time when we come to the access control clauses.

#7 - 03/27/2013 02:31 PM - Redmine Admin

Original Bugzilla ID was 5696