

Infrastructure - Task #3139

Story # 3136 (Closed): Inconsistent data in production 1.0.3 release

Identify and correct missing identifiers table entries on ORC and UNM (w/xml_documents)

2012-08-15 16:13 - Robert Waltz

Status:	Closed	Start date:	2012-08-15
Priority:	Immediate	Due date:	
Assignee:	Ben Leinfelder	% Done:	100%
Category:	Environment.Production	Estimated time:	0.00 hour
Target version:	Sprint-2012.29-Block.4.3	Story Points:	
Milestone:	CCI-1.0.4		
Product Version:	*		

Description

The xml_documents table entries are missing identifiers table entries. If there is an entry in the xml_documents table, then there should be an entry in the identifiers table crossreferencing it to the systemMetadata table.

Use the following query on the 3 production machines:

```
su postgres -c "psql -d metacat -c \"select count(*) from xml_documents x \
where not exists (select * from identifier i where x.docid = i.docid and x.rev = i.rev)\""
```

It will result in 5 results on ORC and 2 results on UNM.

History

#1 - 2012-08-15 17:20 - Ben Leinfelder

- Status changed from New to In Progress

```
/**
 * for https://redmine.dataone.org/issues/3139
 */
-- on ORC (missing the most, UNM is missing only two already included in the 5)
select docid, rev from xml_documents x where not exists (select * from identifier i where x.docid = i.docid and x.rev = i.rev);

-- on UCSB, find the missing mappings
select guid, docid, rev
from identifier
where docid in
('autogen.2012081123435940734',
'autogen.2012081201561580946',
'autogen.2012081205115812190',
'autogen.2012071013471468072',
'autogen.2012071014032840973')
order by docid;

-- on ORC
BEGIN;
insert into identifier (guid, docid, rev)
values ('ark:/13030/m5959g94/1/cadwsap-s3910015-005.xml', 'autogen.2012081123435940734', '1');
insert into identifier (guid, docid, rev)
values ('ark:/13030/m5cv4gcn/1/cadwsap-s5010028-023.xml', 'autogen.2012081201561580946', '1');
insert into identifier (guid, docid, rev)
values ('ark:/13030/m5hq3xn3/1/cadwsap-s3910005-028.xml', 'autogen.2012081205115812190', '1');
COMMIT;

-- on UNM
BEGIN;
insert into identifier (guid, docid, rev)
values ('ark:/13030/m5cv4gcn/1/cadwsap-s5010028-023.xml', 'autogen.2012081201561580946', '1');
insert into identifier (guid, docid, rev)
values ('ark:/13030/m5hq3xn3/1/cadwsap-s3910005-028.xml', 'autogen.2012081205115812190', '1');
COMMIT;
```

```
-- These are duplicate EML docs on ORC that have no mapping to their pid - a different autogen maps to the pid already.  
select * from xml_documents where docid in ('autogen.2012071013471468072', 'autogen.2012071014032840973');
```

```
/**  
* 'autogen.2012071013471468072' shows 'doi:10.6085/AA/LNDXXX_013MTBD004R00_20030217.50.2' in EML file on ORC, but that PID maps to  
'autogen.2012071023184180742' -- files are the same on disk  
* 'autogen.2012071014032840973' shows 'doi:10.6085/AA/LNDXXX_013MTBD012R00_20010321.50.5' in EML file on ORC, but that PID maps to  
'autogen.2012071023241907218' -- files are the same on disk  
**/
```

```
-- on all three servers this is correct
```

```
select * from identifier where guid in ('doi:10.6085/AA/LNDXXX_013MTBD004R00_20030217.50.2',  
'doi:10.6085/AA/LNDXXX_013MTBD012R00_20010321.50.5');
```

```
--remove the duplicate docs from ORC
```

```
BEGIN;  
DELETE FROM xml_queryresult WHERE docid in ('autogen.2012071013471468072', 'autogen.2012071014032840973');  
DELETE FROM xml_path_index WHERE docid in ('autogen.2012071013471468072', 'autogen.2012071014032840973');  
DELETE FROM xml_index WHERE docid in ('autogen.2012071013471468072', 'autogen.2012071014032840973');  
DELETE FROM xml_documents WHERE docid in ('autogen.2012071013471468072', 'autogen.2012071014032840973');  
DELETE FROM xml_nodes WHERE docid in ('autogen.2012071013471468072', 'autogen.2012071014032840973');  
COMMIT;
```

```
-- check our work
```

```
select docid, rev from xml_documents x where not exists (select * from identifier i where x.docid = i.docid and x.rev = i.rev);
```

#2 - 2012-08-15 18:35 - Ben Leinfelder

- Status changed from In Progress to Closed