# Metacat - Task #6004

Feature # 5810 (Closed): Implement SOLR-based search

## Figure out why there are only 422 documents indexed in mn-demo-4.test.dataone.org

06/17/2013 10:30 AM - Jing Tao

| | | | |
|---|---|---|---|
| **Status:** | Resolved | **Start date:** | 06/17/2013 |
| **Priority:** | Normal | **Due date:** | |
| **Assignee:** | Jing Tao | **% Done:** | 0% |
| **Category:** | | **Estimated time:** | 0.00 hour |
| **Target version:** | 2.1.0 | **Spent time:** | 0.00 hour |

**Description**

There are more 4000 data objects in the mn-demo-4. But we only gets 422 documents indexed.

### History

**#1 - 06/17/2013 10:31 AM - Jing Tao**

*- File solr-index.log added*

Attached a log file for the indexing process.

**#2 - 07/02/2013 12:20 PM - Matt Jones**

*- Tracker changed from Feature to Bug*

**#3 - 07/02/2013 12:38 PM - Jing Tao**

I queried the database and got the following result. There are 4998 in the xml_documents and 0 in the xml_revisions.

Something must be wrong because from the objects list, it has lots of archived documents.

```
metacat=> select count(*) from xml_revisions;
count
-------
0
(1 row)

metacat=> select count(*) from xml_documents;
count
-------
4998
(1 row)
```

**#4 - 07/02/2013 12:51 PM - Jing Tao**

*- Assignee set to Jing Tao*

**#5 - 07/09/2013 03:41 PM - Jing Tao**

After we modified the SystemMetadataEventListener class to listen to a dedicated map for indexing, I indexed the metacat again. Now we got 1024 documents. This is a reasonable number. Since the documents' revision in this metadata were messed up, we don't know the exact number which needs be indexed.

I will index the server again to see the result.

**#6 - 07/10/2013 01:09 PM - Jing Tao**

I did couple queries and found the archived value in the systemmetadata was messed up. Please see the bug: https://projects.ecoinformatics.org/ecoinfo/issues/6035

So i did this query on the postgresql:

```
metacat=> select count(*) from systemmetadata where obsoleted_by is null and archived=false;
count
-------
340
(1 row)
```

It seems there should be 340 documents being indexed.

However, our number is much bigger than this.

Here is the reason i think why:

If we have two documents foo.1.1 and foo.1.2.

foo.1.1 should be archived by foo.1.2.

If the systemmetadata information is correct, there is only foo.1.2 being indexed.

However, in our case, foo.1.1 was marked to be obsoleted_by foo.1.2 and the archived value is false.

If the indexing order is foo.1.1 first, then foo.1.2, the solr index will only have one doc - foo.1.2.
If the indexing order is foo.1.2 first, then foo.1.1, the solr index will have both foo.1.1 and foo.1.2.

This is the reason why we have more documents thank it should be.

I was thinking the sort the index queue by the systemmetadata modification time. However, it is not guarantee that the foo.1.1 has earlier systemmetadata modification time than foo.1.2.

**#7 - 07/12/2013 03:58 PM - Jing Tao**

So the number of the total documents which should be indexed is 340. I tested and got 337. And there is no error in the index_event table.
The three which were not initially indexed were indexed by the reindex of the metacat API.

Now I will switch from the embedded solr server to http solr server to index it.

Also the index in the dev2 machine is under way.

**#8 - 07/12/2013 05:25 PM - Jing Tao**

Using the http solr server, the Metacat-index still indexed 337 documents which is less than 340. The 3 documents were not indexed are:
doi:10.5072/FK2/LTER/knb-lter-gce.333.5
doi:10.5072/FK2/LTER/knb-lter-gce.331.12
doi:10.5072/FK2/LTER/knb-lter-gce.335.6

They are the same ones when the embedded server did the index.

I looked the list of 340 documents which should be indexed, it has:

doi:10.5072/FK2/LTER/knb-lter-gce.333.6
doi:10.5072/FK2/LTER/knb-lter-gce.333.5

doi:10.5072/FK2/LTER/knb-lter-gce.331.13
doi:10.5072/FK2/LTER/knb-lter-gce.331.12

doi:10.5072/FK2/LTER/knb-lter-gce.335.6
doi:10.5072/FK2/LTER/knb-lter-gce.335.7

Those 3 documents were not indexed apparently are archived versions.

Aha, i see the reason. The systemmetadata table are messed up.

The archived value for doi:10.5072/FK2/LTER/knb-lter-gce.335.6 is false and the value of obsoleted_by is null. So it is not an archived document and we decided put it into the index queue.
But the obsoletes value for doi:10.5072/FK2/LTER/knb-lter-gce.335.7 is doi:10.5072/FK2/LTER/knb-lter-gce.335.6. This means doi:10.5072/FK2/LTER/knb-lter-gce.335.6 is an archived version. It is a contradiction! When we indexed doi:10.5072/FK2/LTER/knb-lter-gce.335.7 , the solr doc for doi:10.5072/FK2/LTER/knb-lter-gce.335.6 will be deleted.

This is the reason why we have 337 documents indexed, which is 3 less than 340.

Same thing to

**#9 - 07/18/2013 08:39 AM - Jing Tao**

*- translation missing: en.field_remaining_hours set to 0.0*

*- Status changed from New to Resolved*

I tried 3 times for indexing by the EmbeddedSolrServer, the results always showed there are 337 documents, which is correct, were indexed.
The index took about 50 minutes.

Close this bug.

## Files

|  |  |  |  |
|---|---|---|---|
| solr-index.log | 1.51 MB | 06/17/2013 | Jing Tao |