Infrastructure - Story #8869

Equivalent identities show owning different amount of packages.

2020-09-16 20:42 - Jing Tao

Status:	New	**Start date:**	2020-09-16
Priority:	Normal	**Due date:**	
Assignee:	Jing Tao	**% Done:**	0%
Category:	d1_indexer	**Estimated time:**	0.00 hour
Target version:			
Story Points:			

Description

Matt reported he could see 188 packages listed as My Dataset on the profile page of KNB, which including two packages he created as his ORCID, when he logged in by his ldap account. However, he only could see two packages when he logged in by his ORCID even though the two identities are set to be equivalent:
https://cn.dataone.org/cn/v2/accounts/http%3A%2F%2Forcid.org%2F0000-0003-0077-4738

It seems the equivalent identities are not commutative.

Matt also noticed that the equivalent ldap id for his ORCID is UID=jones,O=NCEAS,DC=ecoinformatics,DC=org, which uses UID rather than uid. He suspected this is the issue to cause the problem.

History

#1 - 2020-09-18 00:52 - Jing Tao

It turns out My Packages sends something in the query like this:

```
rightsHolder:(
http\://orcid.org/0000-0003-0077-4738 OR
"CN=Matt Jones A729,O=Google,C=US,DC=cilogon,DC=org" OR
UID=jones,O=NCEAS,DC=ecoinformatics,DC=org
) OR writePermission:(
http\://orcid.org/0000-0003-0077-4738 OR
"CN=Matt Jones A729,O=Google,C=US,DC=cilogon,DC=org" OR
UID=jones,O=NCEAS,DC=ecoinformatics,DC=org
) OR changePermission:(
http\://orcid.org/0000-0003-0077-4738 OR
"CN=Matt Jones A729,O=Google,C=US,DC=cilogon,DC=org" OR
UID=jones,O=NCEAS,DC=ecoinformatics,DC=org
)
```

Those identities starting with the uppercase UID are from the dataone identity service. The format is normalized to comply with RFC2253
However, the rights_holder solr index stored in the Solr index is the lowercase uid=jones,o=NCEAS,dc=ecoinformatics,dc=org and the solr query uses the exactly match. So Solr thinks they are different subjects and wouldn't return the correct result.
I believe the solr index access control has the same issue.

At the dev meeting on September 17, we thought there are probably three solutions:

1. The identity service response for the subject info will return both the lowercase and uppercase equivalent ldap accounts. The drop back will be the response will look messier and it can't handle the white spaces and order issue in the ldap account.

2. In the solr server, the fields such as rightsHolder, readPermission, writePermission and changePermission will be case insensitive. I tried this approach and it worked. The drop back will be to change the index schema, to reindex everything and still not to handler the white spaces and order issue in ldap accounts.

3. In d1_index_processor, value of fields readPermission, writePermission and changePermission will be normalized to comply with RFC2253 before store them. So those values stored in the solr index will exactly match the ones from the dataone identity service whose ldap accounts already comply with RFC2253. The drop back is we still at least reindex partial objects.

In the cn, I did a query to look how many objects we have, whose rights holder and access principle start with lowercase cn or uid. I got about 678,177.

```
select count(distinct systemmetadata.guid) from systemmetadata join xml_access on xml_access.guid=systemmetadata.guid where rights_holder like 'cn%' or rights_holder like 'uid%' or principal_name like 'cn%' or principal_name like 'uid%'
```

Any thoughts and comments will be appreciated.

#2 - 2020-09-24 22:33 - Jing Tao

- Category changed from d1_portal to d1_indexer

#3 - 2020-09-24 22:34 - Jing Tao

Create a ticket for a temporary fix in the client MetacatUI:
https://github.com/NCEAS/metacatui/issues/1531

However, we still need to fix it through the solution 3 (normalize the ldap account in the index processor).