Infrastructure - Task #2058

Story # 2056 (Closed): Metacat upgrade procedure

Generate ORE map for Metacat objects

2011-11-28 18:13 - Ben Leinfelder

Status:	Closed	**Start date:**	2011-11-28
Priority:	Normal	**Due date:**	
Assignee:	Ben Leinfelder	**% Done:**	100%
Category:	Metacat	**Estimated time:**	0.00 hour
Target version:	Sprint-2012.05-Block.1.3		
Milestone:	CCI-1.0.0	**Story Points:**	
Product Version:	*		

Description

When we have datapackages in Metacat (e.g. EML metadata) we should generate ORE documents during the Metadata/DataONE upgrade.

History

#1 - 2011-11-29 00:24 - Ben Leinfelder

Since ORE maps are just objects floating around in DataONE/Metacat, they each have SystemMetadata to describe them. For these auto-generated ORE maps, what should the access control rules be on the ORE object? We could have them completely readable by anyone, or I suppose they could use the same access control rules that are defined for the objects they are packaging. But which rules for which object? If it's a single EML file with some data files, it would probably make sense to use the EML-defined access control rules. For other object formats we'd defer to any access control rules defined in the system metadata for the "organizing" metadata object.
For Metacat I know we are focusing on EML 2.x.x for ORE map generation. What about FGDC packages? Are there other objects in Metacat that need ORE maps generated?
Can we have more than one ORE map for a given data package? How should we resolve differences if there are multiple? (I'm worried about creating an ORE map, then accidentally creating another one very similar to it again).

#2 - 2011-11-29 00:52 - Dave Vieglais

ORE maps should have the same visibility as the most restrictive rules associated with the science metadata document(s) they reference. So for the typical case of one science metadata document describing one to several data objects, the ORE object will have the same permissions as the science metadata document.

ORE maps are not tied to a specific type of science metadata, so should be generated for each data package regardless of how the science metadata is expressed. For the KNB, it makes sense to prioritize ORE generation for EML, followed by FGDC and subsequently less frequently used metadata formats.

Technically, there's nothing to prevent multiple ORE maps for a single data package, however it is something that we should try to avoid especially since there's no way to delete duplicate objects from DataONE. From DataONE's perspective two ORE maps would be conceptually identical (and thus duplicates) if they referenced the same objects with the same predicates.

#3 - 2011-12-17 00:48 - Ben Leinfelder

Additional notes are being kept about this in Metacat's bugzilla:

http://bugzilla.ecoinformatics.org/show_bug.cgi?id=5522
and in an etherpad:
http://epad.dataone.org/20111215-knb-upgrade-process

The general feeling is that we can generate ORE maps for ALL KNB packages. When other nodes that had previously been replicating with the KNB via the Metacat replication infrastructure come online, they can generate ORE maps for any packages that *do not already have ORE maps* -- this will require a look up on the CN for matching/equivalent ORE maps for a given data package. Once the other MNs are online, we can reassign those first ORE maps to them using the SystemMetadata.authorativeNode attribute.

#4 - 2011-12-17 00:49 - Ben Leinfelder

- % Done changed from 0 to 70

- Category set to Metacat

- Assignee set to Ben Leinfelder

The code is in place in Metacat, except for the CN lookup for existing ORE maps.

#5 - 2012-01-19 01:11 - Ben Leinfelder

- Status changed from New to Closed

the details of this are being managed in Metacat - the code is there, but the exact timing will be determined in individual metacat deployments