Metacat - Bug #5522

download linked KNB data and convert links in EML to ORE packages

10/28/2011 02:06 PM - Matt Jones

Status:	New	**Start date:**	10/28/2011
Priority:	Normal	**Due date:**	
Assignee:	ben leinfelder	**% Done:**	0%
Category:	metacat	**Estimated time:**	0.00 hour
Target version:	2.x.y	**Spent time:**	0.00 hour
Bugzilla-Id:	5522		

Description

The KNB data sets, and EML data in general, represent linkages to data as online/url linkages in EML documents. When we convert to the KNB to a DataONE Member Node, we need a mechanism to convert these EML packages to create DataONE ORE-base data packages. Depending on the specific situation, different steps will need to be taken:

1) For packages that arrive via the DataONE services, do nothing
2) For packages that arrive via the Metacat and EcoGrid services, check all online/url links:
a) if it is an ecogrid:// link, then create the corresponding link in an ore document
b) if it is a URL marked as "information" in EML, ignore it
c) if it is a URL marked as "download" in EML, then:
i) attempt to download the data, and if successful
-- check if it is real data (hard to do, but filtering out obvious HTML errors, login pages, HTML pages, etc would be tractable)
-- insert it into the MN using the permissions and policies specified in the EML document (need to determine what the ID would be for this object -- maybe the original URL, but need to ensure uniqueness and < 800 chars, etc)
-- add a link to the ORE document for this dataset
d) insert the final ORE document that's been assembled (need to determine the identifier to use)

This utility method should be callable in two ways:
1) For an existing EML document already in metacat, likely to be run on initial conversion and periodically to be sure all proper data packages are created
-- need to be sure that this doesn't create duplicate packages
2) On any INSERT or UPDATE calls
-- when EML is updated, need to rebuild the package
-- when data objects are updated, need to rebuild the package
-- but need to watch out for sequential ops not interfering (e.g., when Morpho updates a data file, then updates a EML file to point at the new data file in a second step, we should only create one new ORE package version)
-- on update calls, be sure to set appropriate obseletes/obsoletedBy properties on the ORE package (the update() calls themselves should handle these properties for the sysmeta for EML and data objects already)

History

#1 - 10/28/2011 02:23 PM - Chris Jones

Creation of SystemMetadata when no SystemMetadata is provided (i.e. any object creation not through the DataONE API) is currently being done in two separate classes: MetacatHandler (on insert and update events) and MetacatPopulator (Ben updated this to work manually for the 0.6.4 API, Robert recently migrated it to the 1.0.0 API).
A class similar to MetacatPopulator should be created that refactors the functionality and those two classes should then call the class with the common code.

#2 - 11/30/2011 01:06 PM - ben leinfelder

Using the Foresite library in Metacat to build the ORE maps fails because of jar dependencies. The dependencies are as follows:
d1_libclient -> Foresite -> Jena -> Xerces
ORE generation fails because Jena is expecting Xerces v2.7 (and works with v2.6) but Metacat was recently upgraded to use v2.11.

Please see the 5291 bug about Xerces 2.11 and sensorML validation. It would be nice to revert back to Xerces 2.7 so that the libraries were compatible, but that might invalidate sensorML documents that are already in Metacat deployments (it was released in Metacat 1.9.5)

#3 - 12/01/2011 12:29 PM - ben leinfelder

Disregard the Xerces panic -- I had an old XercesImpl.jar hanging out in my classpath.

#4 - 12/08/2011 03:52 PM - ben leinfelder

Items I am suspicious about:

2(c)(i) -- Generating new objects from external data (URLs) that metadata points to. There's usually a reason they are not in Metacat, right? Some of them might be very large?

2(b) (the second set) -- If we update only a data object via the Metacat API, nothing else should happen. If it is part of an EML package, the EML file will also be updated (to use the new data object's revision number). So nothing should be triggered by the data update in terms of ORE regeneration.

2(?) -- There's currently only a loose association between ORE documents in Metacat and the documents they describe (which are assumed to be in Metacat but are pointed to with DataONE endpoints).
So if we do update an EML package, we'll have to [somehow] search our local Metacat for any ORE package that uses the EML package as a basis for the ORE package, set it as obsoletedBy the new ORE package we generate for the EML file and add that new ORE file to Metacat. It's that step where we find the ORE files that use our EML file that scares me -- is this just a search against the ORE RDF/XML file? That's as formal as we can get with the current infrastructure or ORE maps in Metacat.

#5 - 12/16/2011 10:03 AM - ben leinfelder

From discussion yesterday:

Converting a node from KNB to D1
---
0. Generate ORE (for all content that doesn't already have it) and sync all KNB content
1. Turn off LTER<->KNB replication
2. Register LTER as MN, sync is off
2a. (LTER avoid generation of ORE for any docs for which that graph already exists -- check CN.search() using SOLR query against ORE maps)
3. On KNB, change Authoritative MN to LTER for all LTER replicas and ORE maps
4. Turn on LTER sync
-- when CN discovers that an object is a replica, CN triggers sysmetaChanged event at LTER, LTER gets latest version of system metadata from CN, KNB's copy is considered a replica.

#6 - 01/05/2012 02:15 PM - ben leinfelder

I'm now downloading remote data that is referenced by EM documents, saving it on the MN with an "autogen" ID and including that in the ORE map.

#7 - 01/13/2012 04:55 PM - ben leinfelder

Additional notes:
1. ORE generation and remote data download need to be done together so that we use the correct URI/identifier for the data object when we generate the ORE map.
2. We should ONLY generate OREs ans download data when they are described by EML hosted on this home server (i.e. not replicas we happen to be hosting).

#8 - 01/24/2013 11:43 AM - ben leinfelder

Are we committed to doing this? LTER was going to be a major source for new data, but perhaps plans have changed. Revisit for 2.1 release.

#9 - 03/27/2013 02:30 PM - Redmine Admin

Original Bugzilla ID was 5522

#10 - 07/09/2013 02:35 PM - ben leinfelder

- Priority changed from Immediate to Normal

- Target version changed from 2.1.0 to 2.x.y