Metacat - Bug #3835

design and implement OAI-PMH compliant harvest subsystem

02/23/2009 06:06 PM - Matt Jones

Status:	In Progress	**Start date:**	02/23/2009
Priority:	Immediate	**Due date:**	
Assignee:	Duane Costa	**% Done:**	0%
Category:	metacat	**Estimated time:**	0.00 hour
Target version:	2.x.y	**Spent time:**	0.00 hour
Bugzilla-Id:	3835		

Description

Metacat's current harvest mechanism works well but is a proprietary system. The Dryad project has proposed to implement an OAI-PMH compliant harvest susbstem for Metacat in order to allow Metacat to interact more effectively with other systems that implement this protocol. This is a tracking bug for the design and implementation of this feature. Other more detailed bugs will be filed for specific tasks. It would be useful if the final system allowed Metacat to act as both an OAI-PMH Data Provider and as an OAI-PMH Service Provider, allowing us to both serve and harvest documents from OAI-PMH servers.

Some issues to consider and discuss:
1) lack of record authorization mechanisms in OAI-PMH. Metacat currently allows harvest with access controls on harvested records. Reverting to a purely OAI-PMH system would eliminate this capability that is used by many of our harvest clients (especially for data, but somewhat for metadata as well). So the design needs to consider a hybrid that allows both public records to be exposed through OAI-PMH and restricted records to be exposed through a protocol like Metacat's that supports access control. What is our design goal here?

2) A corollary of (1) is how to determine who is allowed to update a given record. Does OAI-PMH assume providers always originate from a constant URL endpoint in order to get around authenticating data providers? This is probably not reasonable for even short periods of time (a few years). A number of sites change domain names over short period of times, and the harvester needs to be able to adjust to these changes, update endpoints, and still handle record replacement. Maybe this is a non-issue if PMH allows provider endpoints to be updated.

3) Date-based change detection in OAI-PMH versus GUID-based versioning in metacat. How should these be reconciled? If a PMH harvest occurs every ten days, but a metadata document is revised three times in that interval, does OAI-PMH only get the most recent version? How are the other versions archived and made accessible over time?

4) Data objects. The Metacat harvester allows one to transfer objects of any type, which is used to harvest both metadata objects of various formats (e.g., EML and FGDC) as well as the associated data objects. Each of these objects has their own unique identifier. How would this be handled under OAI-PMH?

A nice background set of slides is here:
http://www.oaforum.org/otherfiles/berl_oai-tutorial_e.ppt

History

#1 - 04/27/2010 03:24 AM - Hannu Saarenmaa

I would be curious to hear what is the status of these developments now? We are very much looking for them.

My view on the issues is to offer a simple solution first and then see if more functionality is really needed. So, use both Metacat and OAI-PMH protocols in parallel. The latter would have a just a public read-only interface, while actions that require authentication or updates would have to be performed via Metacat protocol. For point 3) on change detection, only serve the latest and not intermediate versions.

#2 - 05/13/2010 01:33 PM - Matt Jones

Hi Hannu --
Regarding Comment #1, I've inquired over email about the status of this feature with Duane Costa, who developed it. As far as I know the OAI-PMH is functional in Metacat -- hopefully Duane will fill in the details. I noticed that it is not documented in the Metacat Administrator's Guide, which we will add to our TODO list for the 1.10 release.

#3 - 05/14/2010 08:03 AM - Duane Costa

Hi Matt,

Support for OAI-PMH is included in the Metacat distribution as of version 1.9.2. Configuring Metacat as an OAI-PMH data provider requires that the Metacat administrator support the service by configuring a set of OAI-PMH properties and activating the 'dataProvider' servlet in the web deployment

descriptor file. Support for OAI-PMH harvesting from a remote data repository into Metacat is also provided.

Documentation is provided in file 'docs/dev/oaipmh/MetacatOaipmh.pdf'. In a metacat-dev email dated 4/16/2009 ('Re: [metacat-dev] Proposed Dryad project integration with Metacat'), we had agreed to place design and planning documents in directory 'metacat/docs/dev/oaipmh', but we had not yet addressed the issue of the Metacat Administrator's Guide. I think the thinking at the time was that this was not yet a mature product, but I agree that it's now time to add user documentation to the Administrator's Guide and to officially support this feature in the next Metacat release, especially given the increased level of demand.

The following links demonstrate its use on a Metacat 1.9.2 test instance for which the OAI-PMH service has been configured and deployed:

http://scoria.lternet.edu:8080/knb/dataProvider?verb=Identify

http://scoria.lternet.edu:8080/knb/dataProvider?verb=ListMetadataFormats

http://scoria.lternet.edu:8080/knb/dataProvider?verb=ListIdentifiers&metadataPrefix=oai_dc&from=2001-01-01&until=2010-01-01
http://scoria.lternet.edu:8080/knb/dataProvider?verb=ListIdentifiers&metadataPrefix=eml-2.0.0&from=2001-01-01&until=2010-01-01
http://scoria.lternet.edu:8080/knb/dataProvider?verb=ListIdentifiers&metadataPrefix=eml-2.0.1&from=2001-01-01&until=2010-01-01
http://scoria.lternet.edu:8080/knb/dataProvider?verb=ListIdentifiers&metadataPrefix=eml-2.1.0&from=2001-01-01&until=2010-01-01

http://scoria.lternet.edu:8080/knb/dataProvider?verb=ListRecords&metadataPrefix=oai_dc
http://scoria.lternet.edu:8080/knb/dataProvider?verb=ListRecords&metadataPrefix=eml-2.0.0
http://scoria.lternet.edu:8080/knb/dataProvider?verb=ListRecords&metadataPrefix=eml-2.0.1
http://scoria.lternet.edu:8080/knb/dataProvider?verb=ListRecords&metadataPrefix=eml-2.1.0

http://scoria.lternet.edu:8080/knb/dataProvider?verb=GetRecord&metadataPrefix=oai_dc&identifier=urn:lsid:knb.ecoinformatics.org:knb-lter-gce:26
http://scoria.lternet.edu:8080/knb/dataProvider?verb=GetRecord&metadataPrefix=eml-2.0.0&identifier=urn:lsid:knb.ecoinformatics.org:knb-lter-and:4056
http://scoria.lternet.edu:8080/knb/dataProvider?verb=GetRecord&metadataPrefix=eml-2.0.1&identifier=urn:lsid:knb.ecoinformatics.org:knb-lter-gce:26
http://scoria.lternet.edu:8080/knb/dataProvider?verb=GetRecord&metadataPrefix=eml-2.1.0&identifier=urn:lsid:knb.ecoinformatics.org:knb-lter-mcr:7

http://scoria.lternet.edu:8080/knb/dataProvider?verb=ListSets

The base URL for OAI-PMH harvesting from the data provider in the above examples is 'http://scoria.lternet.edu:8080/knb/dataProvider'. Note that 'eml-2.0.0', 'eml-2.0.1', and 'eml-2.1.0' are individual 'metadataPrefix' parameters, so each requires a separate harvest operation.

Finally, I apologize for not responding to Hannu's query. I will add a copy of this email reply as a comment in Bug #3835 for completeness.

Thanks,
Duane

Matt Jones wrote:

> Hi Duane and Ryan,
>
> I've had several requests from different groups who are interested in using the OAI-PMH harvester that you added to Metacat as part of the Dryad project. I was wondering if you could update me on the status of that feature -- what was developed, what is tested, and what has been documented. I noticed that the bug for this enhancement has not been updated, although it does contain a request from Hannu for a status update: http://bugzilla.ecoinformatics.org/show_bug.cgi?id=3835
>
> Also, I was searching for PMH documentation in the Metacat Administrator's Guide, and was unable to find any. Did you document the feature and its configuration elsewhere? If so, it would be great to add it to the admin guide in a new subsection following the current harvester documentation (which I just updated a few days ago to correct some issues).
>
> I would like to release this as a feature in the next Metacat release (1.10) for the DataONE member node implementation. Do you think that is feasible?
>
> Thanks,
> Matt

#4 - 05/14/2010 11:40 AM - Duane Costa

Matt Jones wrote:

> Thanks, Duane. I had missed that document, which seems quite complete after looking it over. Would you be willing to incorporate the technical details of the documentation into the Administrator's Guide? I think most of what you wrote could go in wholesale as it is now.
>
> The one major issue is that we have tried to make configuration pretty easy for people. Maybe we need to add a new screen to the Metacat administration utility that allows people to turn on and off OAI, and possibly set needed properties? For the most part, the default URLs would be fine, and its ok to enable the servlet by default, so this configuration might be a simple checkbox labeled 'Enable OAI-PMH?'. Do you think that would work?
>
> Matt

Matt,

Yes, with regard to both the Administrator's Guide and the Metacat administration utility, these sound like the best way to go. I'll add this info to the Bugzilla bug (#3835).

I think this works well in terms of scheduling, too. Between May 1 through August 31, I am working on the second year of the Dryad/LTER project. Ryan or Mark will correct me if I'm wrong, but it seems that adding the finishing touches on the Metacat OAI-PMH work that was started last year would be a worthwhile part of this year's effort (although I should add that the main focus this year is to integrate the LTER controlled vocabulary with other vocabularies using the resources provided by a project called HIVE -- Helping Interdisciplinary Vocabulary Engineering).

Duane

#5 - 03/27/2013 02:24 PM - Redmine Admin

Original Bugzilla ID was 3835