Infrastructure - Task #8817

Configure sitemaps on the CN

2019-06-06 23:52 - Bryce Mecum

Status:	New	**Start date:**	2020-02-05
Priority:	Normal	**Due date:**	
Assignee:	Jing Tao	**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:			
Milestone:	None	**Story Points:**	
Product Version:	*		

Description

Support for sitemaps landed last fall in Metacat: https://github.com/NCEAS/metacat/pull/1283. Sitemaps are good for users but especially for search engines and DataONE's Search Catalog could benefit from having sitemaps enabled. A sitemap could help crawlers discover all of the datasets in DataONE. The CNs already run Metacat and should use Metacat's sitemaps ability to generate sitemaps for all content.

To enable sitemaps on the CNs, a few things seem to be needed. I'm not very familiar with how the CNs get built so I may be wrong or be missing things:

- Sitemaps rely on two properties in Metacat's metacat.properties file, which should have values: sitemap.location.base=https://search.dataone.org/ and sitemap.entry.base=https://search.dataone.org/view
- The Apache config the CNs are built with need to serve the sitemap_index.xml and individual sitemaps from the Tomcat webapps dir. Metacat generates sitemaps in the sitemaps subfolder (e.g., /usr/lib/tomcat8/webapps/metacat/sitemaps). A Directory directive should work so long as filesystem permissions are set up for Apache to see the files.
- We need a robots.txt that points at the sitemap index file at search.dataone.org which provides the entrypoint to sitemap_index.xml
- Metacat generates sitemaps with a recurring job mechanism that's internal to Metacat. AFAIK this job isn't turned on when Tomcat loads Metacat and a request has to get sent to the admin API which turns this job on as a side-effect. We might want to change this to reduce maintenance burden or chance of having stale sitemaps

Dave nominated Jing for this work and has targeted this for the next CCI release. I'm not sure which that is so please select whichever one is appropriate.

Note this relates to https://redmine.dataone.org/issues/8693 which we've delayed because Google's crawler infrastructure has changed and DataONE is now visible by Google. Google staff have indicated we only need to send them a robots.txt that points to our sitemaps for them to begin crawling.

Subtasks:

Task # 8858: Update CN Apache configs in version control with directives to support sit...	**New**

History

#1 - 2020-02-26 20:36 - Bryce Mecum

Also oughta set up robots.txt just to be good web people.

#2 - 2020-03-10 18:02 - Bryce Mecum

Ran into a speedbump. The query Metacat uses to find the objects to put in the sitemaps filters the systemmetadata table for objects with a format_id that is in the xml_catalog table's public_id column. This worked on MN deployments because we have things set up to validate all metadata. This isn't the case on the CNs so Metacat on the CNs aren't creating complete sitemaps (they're missing objects of some format IDS).

Need to talk with Jing about this and see if I can find a good solution or at least a workaround.

#3 - 2020-03-10 18:04 - Bryce Mecum

- File xml_catalog.csv added

I attached a cleaned up copy of the xml_catalog table. Notice how the format IDs are split between the public_id and format_id column like gmd-noaa and gmd-pangaea are. It might be possible to just edit the query here to pull in values from either column.

#4 - 2020-03-12 18:30 - Bryce Mecum

Jing and I chatted about this just now. I've written up the details on https://github.com/NCEAS/metacat/issues/1434 but will include them here:

The Sitemap class creates one or more sitemap files with links to the landing page for every publicly-readable, non-obsoleted, metadata object Metacat knows about. It does this by querying the identifier and system_metadata tables and filters to objects matching the above criteria.

When testing on the CNs, we noticed a large chunk of metadata records weren't being included in the sitemaps and my best guess at this point is that it's due to how we find out *which* formats are metadata formats: Querying the public_id column of the xml_catalog table. This worked great for most/all MN installations but not on DataONE for NOAA and PANGAEA ISOTC211 docs because their public_id is set to http://www.isotc211.org/2005/gmd, not http://www.isotc211.org/2005/gmd-{noaa|pangaea} which meant those -noaa and -pangaea docs weren't getting picked up by the Sitemap class.

I ran this by @taojing2002 and he recommended we instead use the ObjectFormatService to find the format IDs which I think is by far the best way to go as it's the most direct source of truth for this list.

I'll plan to refactor the Sitemap class to get this working.

Files

xml_catalog.csv	4.51 KB	2020-03-10	Bryce Mecum