Infrastructure - Task #8858

Task # 8817 (New): Configure sitemaps on the CN

Update CN Apache configs in version control with directives to support sitemaps

2020-02-05 20:02 - Bryce Mecum

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

Description

Sitemaps are located on disk in ${tomcat_webapps_dir}/${context}/sitemaps as sitemap_index.xml and sitemap%d.xml (for each sub-sitemap).

The rule we've come up with is:

RewriteRule ^/(sitemap.+) /metacat/sitemaps/$1 [R=303]

History

#1 - 2020-02-05 20:50 - Bryce Mecum

Did a test right now with sandbox and realized this is trickier than I thought. On sandbox, the CN stack (i.e., metacat) is living on a separate VM from the one running Apache. So a hit to https://search-sandbox.test.dataone.org/sitemap_index.xml needs redirect over to cn-sandbox.test.dataone.org/sitemap_index.xml which will result in a broken sitemap. May have to use proxy the request instead of merely rewriting it as we do on the MNs.

#2 - 2020-02-12 13:43 - Dave Vieglais

If the sitemap entries contain the correct path then ProxyPass and ProxyPassReverse is a simple way to expose the sitemaps.

If the paths need to be adjusted then either correct them on the host or use mod_rewrite for the proxy. Using mod_rewrite will require more resources for the transformation.

An alternative is to rsync them across to search.dataone.org after generation.

Another alternative is to pull them from search.dataone.org and apply a transform on pull.

In all cases, it is necessary to preserve the modified timestamp of the sitemaps so that the correct Last-Modified header is provided with the response. Some harvesters will use the timestamp to determined if further action is required.

Here's a shell script to pull the sitemaps, adjust the URLs and preserve file timestamps:
https://gist.github.com/datadavev/8f2ed113bfa16e017a12e0a27f439e5a

For example, run for cn-stage-ucsb-1: https://search-stage.test.dataone.org/sitemaps/sitemap_index.xml

#3 - 2020-02-12 19:37 - Bryce Mecum

Thanks Dave.

You wrote:

> If the sitemap entries contain the correct path then ProxyPass and ProxyPassReverse is a simple way to expose the sitemaps.

This is the case, so I'll go ahead and move forward with a reverse proxy, do some testing, and update back here.

Your script and approach below is slick so thanks for working that up. May come in handy in the future.

#4 - 2020-02-14 02:04 - Bryce Mecum

Alright, ran a test on STAGE today and this worked nicely.

On MetacatUI host:

```
ProxyPassMatch "^\/(sitemap.+)" "https://cn-stage.test.dataone.org/$1"
ProxyPassReverse "^\/(sitemap.+)" "https://cn-stage.test.dataone.org/$1"
```

On Tomcat / CN Stack host:

```
ProxyPassMatch "^\/(sitemap.+)" ajp://localhost:8009/metacat/sitemaps/$1
```

Note this takes advantage of mod_proxy_ajp and the AJP connector defined in Tomcat's server.xml which I saw already set up and enabled on the CN. It's also my preferred way of running Apache w/ Tomcat these days.

I'm going to coordinate a restart on STAGE with the team sometime soon and discuss the above config and a plan to make the change on search.dataone/cn-ucsb-1.

#5 - 2020-02-24 19:26 - Bryce Mecum

Restart was coordinated last week and things look great on cn-stage.