SEEK - Bug #2174

Problems with multiple eml2datasource actors in a Kepler workflow

09/01/2005 08:51 AM - Dan Higgins

Status:	Resolved	**Start date:**	09/01/2005
Priority:	Normal	**Due date:**	
Assignee:	Matt Jones	**% Done:**	0%
Category:	seek-general	**Estimated time:**	0.00 hour
Target version:	Unspecified	**Spent time:**	0.00 hour
Bugzilla-Id:	2174		

Description

I have recently created several Kepler workflow that contain multiple eml200datasources. In particular, see 'workflows/eco/IPCC_Base_Layers.xml' and 'workflows/eco/GDAL_h1K_NS_Composite.xml'. The first was created by seaching for 'IPCC' and dragging the 10 baseline historical (1961-1990) climate icons to the graph pane and then attaching to ClimateFileProcessor actors to create *.asc files for use in GARP. The second contains 12 Hydro1K datasources (6 for North America and 6 for South America). These data sources are projected, rescaled, and merged into 6 *.asc files with the same resolution as the the IPCC output files.

I have run into several problems with these workflows that are apparently due to the multiple eml200 datasources. These problems include:

1) First of all, the workflows are somewhat awkward to create. One needs to seperately drag each datasource from the search results to the graph frame and then configure the actor. Configuration is somewhat confusing with a very large collection of options. (At a minimum, we need documentation for all the choices somewhere.) Maybe we need some sort of datasource array (or sequence) actor which could represent multiple data sources?

2) When one drags a datasource to the graph pane, the actor turns red while loading data/metadata from the ecogrid and then turns yellow when no longer busy. The delay isn't too bad while building the workflow (and it is in its own thread, so it doesn't stop all other activities). However, when you reopen a workflow with multiple data sources, you have to wait for all the sources to turn from red to yellow. This can take a ridiculously long time on the machine that was not used to create the workflow. [The time may be very, very long the first time because the data has not been cached locally. It took 4-6 HOURS to open the Hydro1K workflow (which has very large files) on a new machine!] And even though each of the sources has its own thread, when there are a number of such actor, they grap so much CPU time that it is not practical to do anything else on a PC.

If I remember correctly, the main reason for the activity when an eml2datasource is dragged to the graph pane is to create the dymanic ports for all the columns in a data table. But the datasources here are not tables and the output ports are determined by the configuration settings. Maybe whatever is going on in the threads when the workflow is opened should be deferred until the workflow is executed? This would let a user open a workflow and examine and edit it without waiting a very long time.

3) There seems to be some errors that occur when trying to open a workflow with multiple data sources. The IPCC workflow will sometimes open OK, but more often all the sources except one will eventually turn from red to yellow. One, however, will remain red and the CPU usage will stay pegged at 100%. The datasource that will not load is not always the same if you shut down java and start again!

Dan Higgins

History

#1 - 11/02/2005 12:09 PM - Dan Higgins

A test for this is the workflows/eco/IPCC_Base_Layers.xml workflow with tries to load 10 IPCC datasource files. This reliably 'hangs' on the last of the datasources (which remains 'red').

#2 - 04/06/2007 01:57 PM - Dan Higgins

Workflows with multiple datasource actors now seem to work OK!

Assume fixed

#3 - 03/27/2013 02:19 PM - Redmine Admin

Original Bugzilla ID was 2174