Kepler - Bug #3671

Configurable workspace directory for holding workflows, data, and run products

11/13/2008 11:04 AM - Timothy McPhillips

Status:	New	**Start date:**	11/13/2008
Priority:	Normal	**Due date:**	
Assignee:	Timothy McPhillips	**% Done:**	0%
Category:	general	**Estimated time:**	0.00 hour
Target version:	3.X.Y	**Spent time:**	0.00 hour
Bugzilla-Id:	3671		

Description

In bug 3558 I requested that a new directory be created on the user's system for each workflow run, and that outputs of the run, trace files, etc, be placed there. In bug 3585 I asked for an API that would make it easy for actors to write output files to this 'run' directory.

But where should these run directories themselves go? I believe we should allow users to specify a directory for holding their 'workspace' in a location of their choosing. In the workspace could go a directory for holding the workflows they develop and use for a particular project (we've done this before in the Kepler/ppod release, but the directory location was fixed), another directory for holding workflow runs, etc.

One alternative would be to hide all this somewhere inside .kepler in the user's home directory. However, I don't think this is the best approach for two reasons. First, the point is to make it easy for users to find their workflows, data, and workflow run products, and to load the latter into other tools for visualization and further analysis. The .kepler directory is hidden and should be used for things that would distract the user if made more prominent. Second, in practice the .kepler directory is frequently deleted (sometimes when installing a new version of Kepler, for example). A user's work should not be deleted at such times, so .kepler should be used only for things that can be regenerated as needed (e.g. data caches).

Another alternative would be to store everything discussed here in a database. However, many workflows (a) generate large numbers of large data files that would be awkward to place in a database, and (b) users often want immediate file-system access to these output files anyway because the other tools they use to review and further analyze their results expect the data to be stored in files. There shouldn't be an extra step of exporting workflow run products from a database to a directory of files after each workflow run in such cases.

I also think users should have the option of creating multiple workspaces, each with their own directories of workflows and runs. A workspace browser in Kepler could make it easy to view workflows or runs from a particular workspace or all of them at once.

Note that all this has ramifications for distributed execution. Following execution on multiple nodes, the files expected to be found in a local run directory will need to be copied automatically from each compute node.

History

#1 - 03/27/2013 02:24 PM - Redmine Admin

Original Bugzilla ID was 3671