Kepler - Bug #3153

Preview problems with EML DataSource (for IPCC data)

02/19/2008 11:44 AM - Dan Higgins

Status:	Resolved	**Start date:**	02/19/2008
Priority:	Normal	**Due date:**	
Assignee:	ben leinfelder	**% Done:**	0%
Category:	actors	**Estimated time:**	0.00 hour
Target version:	1.0.0	**Spent time:**	0.00 hour
Bugzilla-Id:	3153		

Description

As of RC1, there is a 'Preview' menu item on the popup menu for EML DataSource actors. If one tries to preview the data for results of an 'IPCC' search, several problem appear.

For small data sets (like predictions for future climate variables where data may have 96 columns and 48 rows), initial header lines are apparently truncated in the preview display. Since a primary reason for the preview is to help understand the data, the header lines should not be removed in the display

With larger IPCC datasets (e.g. historical data with 720 x 360 data arrays), clicking on the preview menu seems to hang Kepler (at least for several minutes) I had to kill Kepler to continue.

Dan Higgins - Feb 19, 2008

History

#1 - 02/19/2008 11:52 AM - ben leinfelder

i'll take a look at this.
Is there a specific data package that you were looking at? (there are many)

#2 - 02/19/2008 12:51 PM - Dan Higgins

I think the problem occurs for any of the IPCC files; specifically, look at
$KEPLER/DEMOS/ENM/IPCC_Change_MaxTemp.xml which has 2 datasource, one small, the other large

#3 - 02/21/2008 12:26 PM - ben leinfelder

This is a problem due to the record delimiter [not] specified in the EML for the data.
the EMParser defaults to:
recordDelimiter = "\r\n";
the large data set:
http://knb.ecoinformatics.org/knb/metacat/dpennington.22.3/knb
has the data records delimited with just "\n"

Is "\r\n" an acceptable default delimiter to use? Should we:
1. default it to the default delimiter used on the platform that Kepler is running (and parsing) on? (not a solution at all, since it's from the system of whomever uploaded the data)
2. mandate that people be rigorous with the metadata so that it matches the data?

I vote for #2

Suggestions on how to proceed are greatly encouraged....

#4 - 02/21/2008 01:11 PM - Matt Jones

For common problems like this, if the record delim is missing from the metadata, we should try a couple of common ones (\n, \r, \r\n) and see if any work. By work it means that we get the right number of columns per record, etc.

Then we should post a warning to the user that we had to guess the structure of the dataset and that the metadata should be updated to be more accurate.

#5 - 02/21/2008 01:14 PM - Matt Jones

For that matter, even if record delimiter is present in the metadata, if we get a parsing error we might want to try a different record delimiter just to be more accepting of real-world input, and again post a worning that the metadata wasn't consistent with the data so we used a different delimiter.

In addition, we probably need to incorporate the data manager lib into Kepler now that it is more mature, so that all of these sorts of parsing decisions

are incorporated in a single, well-maintained library. This would obviously be a big refactor for Kepler (but was our original internt with the datamanager lib).

#6 - 02/21/2008 02:31 PM - ben leinfelder

We can display the message when the metadata is first parsed. Right now I have it as a MessageHandler warning for *each* entity in the datapackage.
We can relax this to only show once if any of the entities in the datapackage have unspecified record delimiters.
Note that the EML is parsed when you either drag the EML actor from the data search results or just open a workflow with an EML actor on the stage.
So you'll get the message every time you open the workflow or use the that actor with nonspecific metadata.

Also - some of these IPCC datasets are quite large and take a while to be crammed into table cells. Might want to revisit a Preview that is just the content of the data file (I'm not super keen on this idea, however).

#7 - 02/21/2008 02:49 PM - ben leinfelder

I've also hooked up the "Allow lenient data parsing" parameter from the EMLDatasource actor to the parsing behavior so that if that option is checked we then try \n, \r, and \r\n as other possible line endings.

#8 - 02/21/2008 02:56 PM - Dan Higgins

Some additional comments:

It should be noted that the IPCC data that originally started this error is not the more common database type table. It is really a spatial grid with numrows x numcols of data. (actually it is a set of such 2D grids, but that is not too important here). Note that it does NOT have numcols of values on each ascii line. It actually has only 10. So the numrows and numcols are logical values not the physical values of the file format. (The only requirement is the the product of numrows x numcols is the total number of values).

I would suggest that for ascii gis data, the preview should just display the ascii exactly as it is saved in the file (with any of the various linefeeds observed, just as standard text editors do). Kepler should only try to parse the data into a table when it is declared a table in the EML. (I don't think it is for the IPCC data?)

#9 - 02/25/2008 08:43 AM - ben leinfelder

added support for just viewing the contents of the cache file rather than trying to cram it into a table.
also, if the the cache file is an archived format, then the preview will display a list of the files in that archive (depends on the configuration of the eml actor in that the output format must be set to emit the file name list and the extension filter is adhered to if present)

#10 - 03/27/2013 02:22 PM - Redmine Admin

Original Bugzilla ID was 3153