Kepler - Bug #2712

Problem with EML2DataSource with extra cols in csv file

12/31/2006 01:27 PM - Dan Higgins

Status:	Resolved	**Start date:**	12/31/2006
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:	2712		

Description

This problem can be seen by seaching for 'biomass'. First result is "1999 Sevilleta NPP Quadrat Sampling Data". Drag onto canvas and configure Data Output Format to return 'As Column Vector'. If you make a SDF workflow and try to display any of the columns to a Display actor the message

"Metadata sees data has 12columns but actually data has 13columns. Please make sure metadata is correct!"

Most of the rows in the table do have only 12 comma-separated columns. However, there are a few rows that have some additional comma-separated comments AFTER the 12th column value! This apparently causes a parsing failure.

I suggest that the parser should be modified to ignore any addition data beyond the last column. This would allow additional comments to the right of actual data columns. Note that R does this when it parses dataframes. Morpho also will display the data in this dataset without 'choking' on additional data off the right in some rows. (Dan Higgins)

History

#1 - 12/13/2007 01:41 PM - Dan Higgins

Need a parameter to turn 'ignoring extra data on and off' (rather than just having an error)

#2 - 01/24/2008 03:26 PM - ben leinfelder

Well, being lenient with this example dataset works...until we get to a record with 2.5 in a column described as an Integer. This is a different problem, and it's not clear if we should be lenient in this case, too.

#3 - 01/24/2008 04:48 PM - ben leinfelder

There is now a new parameter in the EML Datasource actor for "Allow lenient data parsing"
When it is checked (true), extra columns are ignored by the DelimitedReader if they are not described in the metadata. If left unchecked (false is the default), then an error will be raised if extra data is encountered.

ps: data type mismatches will continue to raise an error - 2.5 is just not an Integer.

#4 - 03/27/2013 02:21 PM - Redmine Admin

Original Bugzilla ID was 2712