Base Language - Bug #5276

resolve issues with multi-byte character processing

04/22/2021 08:37 AM - Greg Shah

Status:	New	**Start date:**	
Priority:	Normal	**Due date:**	
Assignee:		**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:			
billable:	No	**case_num:**	
vendor_id:	GCD	**version:**	

Description

History

#1 - 04/22/2021 08:44 AM - Greg Shah

We need to review our text processing (throughout the FWD runtime) to ensure that multi-byte character processing is correct. Today we use features like String.charAt() and String.length() without any extra consideration, but that is not correct for the surrogate character cases. Our code only works today for the "Basic Multilingual Plane" (BMP) which are the characters that can be encoded in a 16-bit value.

This has 2 aspects:

- We need to check the 4GL behavior of these cases and fix any compatibility issues. Some issues here may come from the difference where the 4GL operates on 8-bit bytes but our internal implementation operates on 16-bit char.
- We need to check for problems that are intrinsic to how Java handles String and char as fixed 16-bit values. In particular, handling surrogate characters which require 2 char to encode will definitely need special attention.

These aspects will overlap but we need to consider both to ensure there is no path through our code which can cause issues.

#2 - 04/05/2023 12:51 PM - Greg Shah

- Assignee set to Joe Davis

#3 - 04/24/2023 10:45 AM - Joe Davis

Status on this: I'm in the middle of converting character.java to use codePointAt instead of charAt where appropriate, and adding unit tests for this case. The unit tests are not intended as a replacement for integration tests written in Progress, just to help codify the expected behaviour during development, while the rest of the MB character support is worked on.

There's a large number of other cases in the source code outside character.java where charAt is used, most of these are going to have to be checked individually to see if it's correct.

#4 - 09/13/2023 08:17 AM - Greg Shah

- Assignee deleted (Joe Davis)