Conversion Tools - Feature #4389

escape support for supplementary unicode characters

11/07/2019 02:33 PM - Greg Shah

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

Description

History

#1 - 11/07/2019 06:52 PM - Greg Shah

Branch 4069a 11374 conversion added support for encoding Basic Multilingual Plane (BMP) unicode characters as represented in the 4GL. BMP chars are the standard UTF-16 (fixed 2 byte representation) characters that are built into the low level Java char and String operations. In the 4GL these are encoded in source using ~uXXXX (Windows and Linux/UNIX) and \uXXXX (only in Linux/UNIX) where X is a hexadecimal digit.

There do exist other unicode characters outside of the BMP which cannot be represented in a single UTF-16 character. These are called supplementary characters and can be encoded naturally in UTF-32. In the 4GL these are encoded in source using ~UXXXXXX (Windows and Linux/UNIX) and \UXXXXXX (only in Linux/UNIX) where X is a hexadecimal digit. The previous 4069a code did not support this form of the escape sequence conversion.

Java can work with these using int types and the Character class to take UTF-32 characters and represent them as two sequential UTF-16 characters (Java char values) which must follow special encoding rules. These are called the high surrogate and low surrogate respectively. These can be encoded in Java source using this form.

Branch 4069a revision 11450 adds conversion support for this supplementary character escape encoding.

Please read the following references for the full details.

Basic Multilingual Plane
Java Lexical Structure
How Java Supplementary Character Encoding Works (really great reference!)
Unicode Escapes in Java
Testing Supplementary Characters
Unicode 12.1 Technical Site