



Main page
Recent changes
Random page
Help

Tools
What links here
Related changes
Special pages
Printable version
Permanent link
Page information

Log in

Page Discussion Read View source View history Search commentsarchive

C 0000004 mmmv string t1

C 1..100

This spec is very likely to change.

The motivation for this x-spec is to have some idea, how to cope with strings in ParaSail.

General Idea

The general idea is that as strings are meant to fit into RAM, they have relatively few characters, regardless of what Unicode codes those characters have. This allows to redefine a runtime-alphabet that is far smaller than the whole Unicode alphabet is and does not contain any unassigned codes and has a maximum index value. Strings and queries are converted to that temporary alphabet before processing and the result is that in stead of 4B or 8B characters probably all string codes fit into 2B, most likely 1B, which makes the processing versions of the strings smaller than their originals might have been and it also normalizes the character memory consumption sizes, which allows to use ordinary byte arrays as the data structure for holding strings. That in turn allows SIMD-instruction based optimizations.

Object Oriented View

The byte array is at the core class and text processing algorithms are implemented as independent classes that are NOT derived from the core class, but that are dependent on the core class. That includes regular expressions, etc.

Conversion from and to Unicode

There should be a separate C++ console application that converts strings from mmmv_string_t1 fomrat to Unicode and vice versa. The console application has some server mode.

This page was last modified on 20 May 2020, at 14:27.
Content is available under [Creative Commons Attribution Share Alike](#) unless otherwise noted.
[Privacy policy](#) [About commentsarchive](#) [Disclaimers](#)

