L2/17-225

Unicode Continuity Characters:

Character Plane Sets and Traversal Characters

Abram Wiebe - abramw@sfu.ca July 14, 2017

Introduction

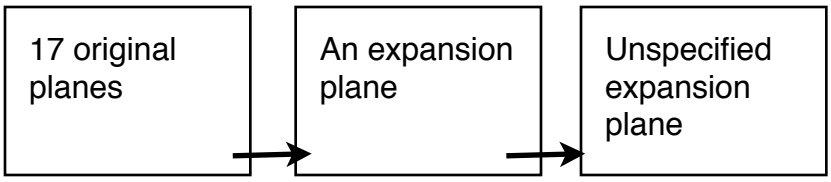
Unicode can currently encode 1,111,998 code points

17 planes × 65,536 code points per plane - 2048 surrogates - 66 non-characters

While it is not unreasonable to expect that this is a sufficient number of characters for all human scripts past an present, it should be noted that as of unicode 9.0 128,237 characters ~11.5% of this space has already been allocated. In order to future proof unicode, if not for now, then for future generations; so that we do not have a repeat of the shortsightedness of ASCII et al. Unicode should have a method to represent other sets of 17 character planes.

Concept

To facilitate the expansion of the character space while remaining compatible with UTF-16 expansions should be encoded in planes having the same format as the original set of 17 planes.



Encoding

A number of special non-printing characters should be assigned to indicate which plane set should be the active plane for the subsequent string of characters. When a unicode string begins the default character plane set shall be the original plane set. Since plane set navigation characters must be specified in every plane set, they should be near the end of the 17th plane.

	Name	Action
0	First Plane Set	Select the original plane as the active plane set(equivalent to goto plane set 0, but short because commonly used)
1	Previous Plane Set	Select the previous plane set as the active plane set
2	Next Plane Set	Select the next plane set as the active plane set
3	Goto Plane Set	Interpret the next 16bit number as the plane set to skip to

Stability of Planes Sets

In order to retain stability across definitions of Unicode:

- an expansion plane cannot be deleted or moved
- asking for the next plane group of the last plane is undefined
- asking for the previous plane of the first plane emits U+FFFF
- every plane set must specify the plane set navigation characters
- every plane set should specify the plane navigation characters in the same location, except that the first plane set may specify them somewhere else.
- the last character of the last plane set should be a character which further extends unicode by putting it in a mode yet to be defined(hopefully for extraterrestrial symbols).

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[1]
Please fill all the sections A, B and C below.
Please read Principles and Procedures Document (P & P) from http://std.dkuug.dk/JTC1/SC2/WG2/docs/principles.html **for guidelines and details before filling this form.**
Please ensure you are using the latest Form from http://std.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html **.**
See also http://std.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html **for latest *Roadmaps*.**

A. Administrative

1. **Title:** *Unicode codepoint expansion by specifing code plane set*
2. Requester's name: *Abram Wiebe*
3. Requester type (Member body/Liaison/Individual contribution): *individual contribution*
4. Submission date: *2017-07-15*
5. Requester's reference (if applicable):
6. Choose one of the following:
   This is a complete proposal: *yes*
   (or) More information will be provided later:

B. Technical – General

1. Choose one of the following:
   a. This proposal is for a new script (set of characters): *yes*
      Proposed name of script: *Expansion*
   b. The proposal is for addition of character(s) to an existing block:
      Name of the existing block:
2. Number of characters in proposal:
3. Proposed category (select one from below - see section 2.2 of P&P document):
   A-Contemporary B.1-Specialized (small collection) B.2-Specialized (large collection)
   C-Major extinct D-Attested extinct E-Minor extinct
   F-Archaic Hieroglyphic or Ideographic G-Obscure or questionable usage symbols *
4. Is a repertoire including character names provided? *Yes*
   a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document? *yes*
   b. Are the character shapes attached in a legible form suitable for review? *n/a*
5. Fonts related:
   a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?
   n/a
   b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):
   n/a
6. References:
   a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? *n/a*
   b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? *n/a*
7. Special encoding issues:
   Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)? *yes*
   Characters are non printing
8. Additional Information:

Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see Unicode Character Database (http://www.unicode.org/reports/tr44/) and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

[1] Form number: N4502-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11, 2005-01, 2005-09, 2005-10, 2007-03, 2008-05, 2009-11, 2011-03, 2012-01)

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before? *no*
   If YES explain
2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)? *no*
   If YES, with whom?
   If YES, available relevant documents:
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? *yes*
   Reference:
4. The context of use for the proposed characters (type of use; common or rare) *Future*
   Reference:
5. Are the proposed characters in current use by the user community? *no*
   If YES, where? Reference:
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? *no*
   If YES, is a rationale provided?
   If YES, reference:
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? *yes*
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? *no*
   If YES, is a rationale for its inclusion provided?
   If YES, reference:
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? *no*
   If YES, is a rationale for its inclusion provided?
   If YES, reference:
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to, or could be confused with, an existing character? *no*
    If YES, is a rationale for its inclusion provided?
    If YES, reference:
11. Does the proposal include use of combining characters and/or use of composite sequences? *yes*
    If YES, is a rationale for such use provided?
    If YES, reference: *The goto plane set character uses the next number as a parameter*
    Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided? *n/a*
    If YES, reference:
12. Does the proposal contain characters with any special properties such as control function or similar semantics? *yes*
    If YES, describe in detail (include attachment if necessary)
    Controls the interpretation of codepoints based on contextual information
13. Does the proposal contain any Ideographic compatibility characters? *no*
    If YES, are the equivalent corresponding unified ideographic characters identified?
    If YES, reference: