Proposal for Encoding Emoji Symbols

L2/09-025R

Date: 2009-Feb-06

Authors:
Markus Scherer, Mark Davis, Kat Momoi, Darick Tong (Google Inc.)
Yasuo Kida, Peter Edberg (Apple Inc.)

This is a proposal to add a set of 674 symbols to Unicode and ISO/IEC 10646. The so-called "Emoji" symbols are in widespread use by DoCoMo, KDDI and Softbank for their mobile phone networks and are interchanged with many other systems. The Emoji symbols proposed here are not yet encoded in Unicode, while the remaining Emoji symbols can be mapped to existing Unicode characters.

Contents

As with the already-approved ARIB symbols, the purpose for adding these symbols is interoperability. The Emoji symbols are encoded in carrier-specific versions of Shift-JIS (as Vendor-Defined Characters), and, in the case of KDDI, in a carrier-specific version of ISO-2022-JP. There are mapping tables in use in the industry between these character sets, with both roundtrip and fallback mappings. These symbols are also supported in web mail services by Yahoo! Mail and Google Mail, and in the Apple iPhone. (Yahoo! Mail and iPhone currently support subsets of Emoji symbols.)

We are taking into consideration the following factors in developing the proposal. These are based upon discussions in the UTC and in symbols subcommittee ad-hoc meetings.

1. **Source separation rule:** If a single carrier separates two characters (anywhere in the character set, so including standard JIS codes), then we mapped them to two separate Unicode characters. (This is a hard and fast rule.)
2. **Reuse:** We mapped to existing Unicode symbols where appropriate. We included unifications with "upcoming" characters in Unicode 5.2 and ISO/IEC 10646 AMD6. In particular, some unifications are with symbols from the ARIB set.
3. **Separating generic symbols:** If Unicode had a set of related symbols, but no one character in the set was as generic as in the Emoji symbol sets, then we encoded a new character. For example, the Emoji sets do not distinguish between waxing and waning crescent moons.
4. **Colors and Animation:** We encoded symbols as characters, abstracting away from colors and animation.
5. **Existing cross-mapping tables:** We followed the tables mentioned above as much as possible, but we tentatively disunified in some cases where the visual images were very different and not semantically associated. For example:
    1. We disunified the 'M' symbol for Metro from the Metro train image. The 'M' symbol would have translation problems. (This is similar to the problems with the international currency symbol and the proposal for a "generic decimal separator".)
    2. On the other hand, we unified the sets of Zodiac symbols, even though the images shown by carriers vary widely. This is because they clearly belong to a cohesive set which corresponds across carriers.
6. **Complete set:** The proposed symbols are designed for complete round-trip conversion of each of the major carriers' symbols, taking unifications into account. This is necessary for interoperability. See the emoji_sources.txt file (L2/09-078) for complete source information.

Note: We tried to avoid disunification in Unicode where there are round-trip mappings between carriers. However, where necessary, the disunification can be done. As the following diagram illustrates, roundtrip mappings between carrier Shift-JIS character sets can be maintained, by having the mapping tables between Unicode and each carrier's Shift-JIS version use appropriate fallback mappings.

KDDI		Unicode		Softbank
x	↔	X	→	y
x	←	Y	↔	y
x	↔			y

Symbol Identifiers

During proposal development, we used stable, internal identifiers like e-02A for the ALARM CLOCK symbol. These are used only for reference during development.

Code point assignments

Some symbols that are closely related to existing ones are allocated in the same blocks, or in related "supplementary" blocks, if there are unassigned code points available there. Most of the symbols are proposed to be assigned code points in a new block on the SMP. No new BMP blocks are proposed for Emoji symbols. Of the 674 symbols, 9 are proposed for encoding on the BMP, and the remaining 665 are proposed for encoding on the SMP.

Properties

Most symbols are proposed with standard symbols properties, with Bidi_Mirroring=False, for example like

```
2702;BLACK SCISSORS;So;0;ON;;;;;N;;;;;
```

Exceptions:

- One symbol, e-B08 LOOPED LENGTH MARK, is proposed as a punctuation character (gc=Pd). (This is related to U+3030 WAVY DASH which also has gc=Pd.)
- There are several symbols with compatibility decompositions. These decompositions are noted in the charts. The set of characters with decompositions follows established practice in Unicode/ISO 10646.

Collation

Default collation order (DUCET/ISO 14651) should follow the example of Dingbats, except:

- e-B08 LOOPED LENGTH MARK should sort together with U+3030 WAVY DASH.
- Enclosed characters with decompositions should sort accordingly, as usual: With tertiary differences from their normalized equivalents.

Proposal Summary Form

Here is a draft proposal summary form for submission to ISO/IEC.

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646
Please fill all the sections A, B and C below.
Please read Principles and Procedures Document (P & P) from
http://www.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://www.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html.
See also http://www.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html for latest *Roadmaps*.
Form number: N3452-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)

A. Administrative

1. **Title:** ***Proposal for Encoding Emoji Symbols***
2. Requester's name: ***Markus Scherer, Google Inc.***
3. Requester type (Member body/Liaison/Individual contribution): ***Individual contribution***
4. Submission date: ***2009-feb-04***
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: ***No***

B. Technical - General

1. Choose one of the following:
a. This proposal is for a new script (set of characters): ***No***
Proposed name of script:
b. The proposal is for addition of character(s) to an existing block: ***Yes***
Name of the existing block: ***Several, see details***
2. Number of characters in proposal: ***674***
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) ***x***
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" ***Yes***
b. Are the character shapes attached in a legible form suitable for review? ***Yes***
5. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard? ***Apple***
If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used:
6. References:
a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? ***Yes***
b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? ***No***
7. Special encoding issue
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)? ***No***

8. 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

athttp://www.unicode.org for such information on other scripts. Also see http://www.unicode.org/Public/UNIDATA/UCD.htmland associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

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.)? ***Yes***

If YES, available relevant documents: ***Search engine and email/chat vendors are involved***

3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? ***Originally Japan***

Reference: ***Vendor-specific subsets of these symbols are available to all Japanese cell phone users***

4. The context of use for the proposed characters type of use; common or rare) ***common***

Reference:

5. Are the proposed characters in current use by the user community? ***Yes***

If YES, where? Reference: ***Japanese cell phone networks, Google Talk, Google Mail***

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)? ***No***

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 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? ***No***

If YES, is a rationale for such use provided?

If Yes, reference:

Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?

If Yes, reference:

12. Does the proposal contain characters with any special properties such as control function or similar semantics? ***No***

If YES, describe in detail (include attachment if necessary)

13. Does the proposal contain any Ideographic compatibility character(s)? ***No***

If YES, is the equivalent corresponding unified ideographic character(s) identified?

If Yes, reference: