

# 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.

**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.

- 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.)
- 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.
- 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.
- Colors and Animation:** We encoded symbols as characters, abstracting away from colors and animation.
- 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:
  - 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".)
  - 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.
- 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

### Contents

- [1 Symbol Identifiers](#)
  - [1.1 Code point assignments](#)
  - [1.2 Properties](#)
  - [1.3 Collation](#)
- [2 Proposal Summary Form](#)



at <http://www.unicode.org> for such information on other scripts. Also see <http://www.unicode.org/Public/UNIDATA/UCD.html> and 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? If YES explain	<i>No</i>
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.)? If YES, available relevant documents:	<i>Yes</i> <i>Search engine and email/chat vendors are involved</i>
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? Reference:	<i>Originally Japan</i> <i>Vendor-specific subsets of these symbols are available to all Japanese cell phone users</i>
4. The context of use for the proposed characters type of use; common or rare? Reference:	<i>common</i>
5. Are the proposed characters in current use by the user community? If YES, where? Reference:	<i>Yes</i> <i>Japanese cell phone networks, Google Talk, Google Mail</i>
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? If YES, is a rationale provided? If Yes, reference:	<i>No</i>
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)?	<i>No</i>
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? If YES, is a rationale for its inclusion provided? If Yes, reference:	<i>No</i>
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? If YES, is a rationale for its inclusion provided? If Yes, reference:	<i>No</i>
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character? If YES, is a rationale for its inclusion provided? If Yes, reference:	<i>No</i>
11. Does the proposal include use of combining characters and/or use of composite sequences? 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:	<i>No</i>
12. Does the proposal contain characters with any special properties such as control function or similar semantics? If YES, describe in detail (include attachment if necessary)	<i>No</i>
13. Does the proposal contain any Ideographic compatibility character(s)? If YES, is the equivalent corresponding unified ideographic character(s) identified? If Yes, reference:	<i>No</i>