Proposal for Encoding Emoji Symbols

N3582
L2/09-025R2

Date: 2009-Mar-05
Authors:
Markus Scherer, Mark Davis, Kat Momoi, Darick Tong (Google Inc.)
Yasuo Kida, Peter Edberg (Apple Inc.)

Proposal

The Unicode Consortium proposes adding 674 UCS characters to enable complete UCS representation of the fixed set of Emoji that have been encoded in carrier-specific extended versions of Shift-JIS or ISO-2022-JP by the three primary cell phone carriers in Japan: NTT DoCoMo, KDDI, and Softbank. These characters are interchanged with many other systems. The Emoji symbols proposed here are not yet encoded in the UCS, while the remaining 114 Emoji symbols are represented by existing UCS characters.

Documents:

- N3582 = L2/09-025R2: This proposal document
- N3583 = L2/09-026R: Chart of proposed characters
- N3585 = L2/09-078: Sources File

Contents

Rationale

As with the already-approved ARIB symbols, the purpose for adding these symbols is interoperability. Interoperability is important for those cellphone carriers and their clients, but not only for them. What is even more important is that the data be handled without loss or corruption by a vast array of other interoperating systems that use the UCS: email systems, search engines, publishing systems, databases, and so on. These symbols are also supported in web mail services by Yahoo! Mail and Google Mail, and in the Apple iPhone.

Users treat Emoji as text and expect that they can be interchanged like any other element of text. There are mapping tables in use in the industry between these character sets, with both roundtrip and fallback mappings. However, the only UCS representation is via Private-Use characters which are subject to misinterpretation and data corruption.

This core set of Emoji was encoded as extensions to Shift-JIS / ISO-2022-JP following a history of similar vendor-specific extensions by other companies and for other purposes in Japan. (For example, the Japanese recording industry standard RIS-506-1996 specifies an extension of Shift-JIS for use in Music CD text, and includes a number of characters similar to the Emoji in the current proposal.) This approach has enabled these Emoji to be used in plain text contexts such as SMS text messages, e-mail subject lines and address book entries, and many users depend on being able to use Emoji from the core set in this way.

For Emoji beyond this core set, vendors have added rich text support, and use approaches such as embedded graphics. Similar techniques (embedded graphics or escaped tags designating Emoji) are also used for emoji support in China and the Republic of Korea, and there is no demonstrated need for encoding of Emoji beyond the proposed core set. The core set cannot be further extended to support additional Emoji without creating interoperability problems even within the cell phone carriers' own networks. However, the existing core set of Emoji will continue to be represented as encoded characters, because users depend on capabilities that result from this and because this core set is already widespread in existing encoded data.

As of December 2008, there are 110.4 million cell phone users in Japan (about 87% of the population), and about 90.6% of the cell phones are G3-enabled for internet use. Emoji are widely used, especially by people under 30. However, a June 2007 survey of 13,000 users — 80% of whom were 30 or older — found that even among this older group, 78% "often" or "sometimes" used Emoji in emails. Respondents reported using a wide variety of Emoji, including Emoji for faces, emotions, weather, vehicles and buildings, food and drink, animals, etc. Especially among younger users, email is mostly or exclusively used on cell phones instead of computers. Among cell phone users, 90% use email primarily on cell phones, and 60% use email exclusively on cell phones. Emoji have been used on Japanese cell phones for 10 years, and there is no evidence that use of Emoji is decreasing.

The UCS encoding of the proposed core set of Emoji is intended for interoperability with existing data generated by Japanese cell phone users. This interoperability requirement is analogous to the rationale for encoding Dingbats, for example. Because the overriding requirement is interoperability, the identity of the proposed characters is determined to some extent by their mappings to and from the vendor-specific encodings, and therefore this proposal includes a Sources table. Establishing character identity by mapping to sources follows the model used for East Asian ideographs.

Note that encoding of the complete set is essential in order to satisfy the interoperability requirement, otherwise data is lost in round-trip conversion and data interchange.

Sources File

The Unicode Consortium requests to make the emoji_sources.txt file (N3585=L2/09-078) a normative part of the contents of ISO/IEC 10646. This can be done by reference.

References

- http://analytica1st.com/analytica1st/index.html
- http://wirelesswatch.jp/2008/04/25/sharp-maintains-top-spot-in-2007
- http://en.wikipedia.org/wiki/Japanese_mobile_phone_culture
- http://whatjapanthinks.com/2007/07/12/japanese-cell-phone-emoji-graphical-icon-usage/
- http://whatjapanthinks.com/2009/02/20/mobile-email-and-emoticons-emoji-and-friends/
- http://whatjapanthinks.com/2009/02/09/why-use-japanese-emoticons/
- http://whatjapanthinks.com/2008/06/26/emoji-versus-kaomoji-graphical-icons-versus-text-emoticons/
- http://mobilemarketing.jp/pressrelease/20080703.html
- http://ip.tosp.co.jp/index.asp
- http://www.nytimes.com/2008/01/20/world/asia/20japan.html
- http://www.newyorker.com/reporting/2008/12/22/081222fa_fact_goodyear
- http://de-view.net/index.php?PN=BVLlXDYd&LU=novel/novel_disp_book&novel_number=22764
- http://www.nttdocomo.co.jp/service/imode/make/content/pictograph/basic/
- http://www.au.kddi.com/ezfactory/tec/spec/emoji_download.html
- http://mb.softbank.jp/mb/service/3G/mail/pictogram/list.html

Factors

The following factors were considered in developing the proposal.

1. **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 (N3585=L2/09-078) for complete source information.
2. **Source separation rule:** If a single carrier separates two characters (anywhere in the character set, so including standard JIS codes), then they are mapped to two separate UCS characters. Similar to the case of CJK Unified Ideographs, this is a hard and fast rule.
3. **Reuse:** Existing UCS symbols are used where appropriate. This includes unifications with "upcoming" characters in ISO/IEC 10646 AMD6 and Unicode 5.2. In particular, some unifications are with symbols from the ARIB set.
4. **Separating generic symbols:** If the UCS had a set of related symbols, but no one character in the set was as generic as in the Emoji symbol sets, then a new character is added. For example, the Emoji sets do not distinguish between waxing and waning crescent moons.
5. **Colors and Animation:** Colors and animation are not part of the encoding. The characters may have associated colors or animation in particular implementations but that is not part of their identity.
6. **Existing cross-mapping tables:** The mapping tables mentioned above were followed as much as possible. For example, the sets of Zodiac symbols are unified, even though the images shown by carriers vary widely. This is because they clearly belong to a cohesive set which corresponds across carriers and is mapped across carriers. Characters were disunified in a small number of cases, where the visual images were very different and not semantically associated. For example, the 'M' symbol for Metro is disunified from the Metro train image. Round-trip 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.

Symbol Identifiers

During proposal development, stable, internal identifiers were used, for example 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.

Special, rarely used, carrier-specific symbols are proposed for encoding in the Emoji compatibility symbols block. They are needed to complete the set for interoperability but are only identified by their source mappings (N3585), not specific glyphs and names.

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 chart (N3583). The set of characters with decompositions follows established practice in Unicode/ISO 10646.

Collation

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

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

Proposal Summary Form

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-Mar-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?	***Yes (as links: see reference links above)***
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: