

絵文字シンボル符号化の提案(日本語訳)

訳注:この翻訳文は ISO/WG2 に Unicode コンソーシアムより提出された英文の原文の内容をより多くの人に知って頂くことを目的に日本語に訳したものです。従ってこれは正式な提案文書ではありません。あくまで便宜上用意されたものです。

(訳: 桃井勝彦, Google Inc.)

N3582

L2/09-025R2

日付: 2009 年 3 月 5 日

提案者:

Markus Scherer, Mark Davis, 桃井勝彦, Darick Tong (Google Inc.)
木田泰夫, Peter Edberg (Apple Inc.)

目次

1. 提案
2. 提案の理由
3. ソースの文書
4. 参考文献
5. 考慮の要素
6. シンボルの識別名
7. コードポイントの割り振り
8. Unicode プロパティ
9. 文字の照合順序
10. 提案要約文書

提案

Unicode コンソーシアムは日本の主要携帯電話3社(NTT docomo, KDDI/AU, そして SoftBank)が現在まで Shift_JIS や ISO-2022-JP の拡張領域にキャリア独自の符号化をして来た固定数の絵文字セットを UCS (ISO/IEC 10646: [国際符号化文字集合](#))で完全に網羅する為に、必要と思われる674の文字を新しく追加することを提案します。

これらの文字は既に色々なシステム間で交換されています。ここで提案された絵文字シンボルは未だに UCS に符号化されていませんが、固定数の絵文字セットの内残りの114の絵文字シンボルは既に UCS の現存の文字で表現できます。

提出文書:

- N3582 = L2/09-025R2: 本提案文書
- N3583 = L2/09-026R: 提案に含まれる文字の表
- N3585 = L2/09-078: ソースファイル (Unicode 文字と各キャリアの Shift_JIS コード対応表)

提案の理由

既に UCS で承認されている ARIB シンボルでも分かるように、これらのシンボルを新規追加する目的はインターオペラビリティです。勿論インターオペラビリティは携帯各社とそのユーザーにとって重要なものですが、それだけではありません。更に重要なことは文字データが、UCS を使用して連携されている多種多様な他のシステム間で、データ・ロスも文字化けもなしに、交換出来るということです。例えば、メールのシステム、サーチ・エンジン、出版システム、データベース、等々。ここで提案されているシンボルは Yahoo! メールや Google メール (Gmail) のウェブメール・サービスで既にサポートされています。更に、Apple 社の iPhone でもサポートされています。

ユーザーは絵文字をテキストとして見なしており、他のテキスト要素と同様に相互交換できることを期待しています。日本の携帯業界では各社の絵文字セット間のクロス・マッピング表が使われており、それにはラウンド・トリップが可能な文字やフォールバック(代替文字)の マッピングが含まれています。しかし UCS を使用した表象は現在私用/外字領域の文字のみでされており、誤って解釈されたり、データ化けを起こす可能性があります。

この提案の絵文字のコアセットは、歴史的に日本の他の会社が色々な理由で作成したベンダー拡張特殊文字と類似しており、それらと同じ様に Shift_JIS や ISO-2022-JP の拡張として符号化されてきました。(例えば、社団法人日本レコード協会が制定した「RIS-506-1996: レコード用文字符号」ではミュージック CD のテキスト用に Shift_JIS の拡張領域に文字を定義しており、それには絵文字と似た文字が幾つか含まれています。)このような方法で今まで SMS テキストのメッセージ、電子メールの件名、アドレス帳のエントリー等のプレーンテキストのコンテキストでも絵文字 が使用可能となっており、多くのユーザーがこのようなやり方でコアセットの絵文字使用が出来ることを頼りにしています。

このコアセット以外の絵文字は各社がリッチ・テキストのサポートを追加して、イメージを埋め込むなどのアプローチを使って対処しています。これと似たような方法、つまり埋め込みイメージやエスケープ・タグを使って絵文字を表現することは中国や韓国でも絵文字をサポートするのに使われており、ここに提案しているコアセット以外の絵文字を符号化しなければならないという必要性はないと思われます。このコアセットを拡張して新しい絵文字をサポートしようとするれば、インターオペラビリティの問題を回避することは携帯会社のネットワーク内でさえ出来ないというのが現状です。しかし、現存する絵文字のコアセットは将来的にも継続して符号化された文字として表現される筈です。何故なら文字としての扱いにより可能なサービスにユーザーは依存しており、このコアセットの絵文字は現存する符号化されたデータにも広く使われているからです。

2008 年 12 月現在日本に於ける携帯ユーザー数は1億千 40 万人(総人口の約 87%) で、その内約 90.6%の携帯電話はインターネット用の 3G 機種です。絵文字は 30 才以下の人たちに広く使われています。しかし 2007 年 6 月に行われた 1 万 3 千人のユーザー(80%が 30 才以上)が回答したアンケートによると、このアンケートの年上の世代でも 78%の人たちが「よく」或いは「時々」絵文字をメールで使うと答えています。回答者によると色々な種類の絵文字を使っており、それには顔を表すもの、感情、天気、車、建物、食べ物、飲み物、動物などを表すものが含まれているとのこと。特に若者の世代では電子メールは殆ど或いは全て、パソコンではなく、携帯電話経由で送受信されています。携帯電話ユーザーの総数の内 90%が主に携帯電話で電子メールを利用しており、60%が携帯電話のみで電子メールを使用しています。絵文字が日本の携帯電話で使われるようになって 10 年になりますが、絵文字の利用が減少しているというような証拠はどこにもありません。

ここで提案された絵文字のコアセットを UCS で符号化することは日本の携帯電話ユーザーが作った現存データとのインターオペラビリティを主目的としています。そしてこのインターオペラビリティの必須条件は例えば Dingbats シンボルの符号化の理由と類似しています。インターオペラビリティが何にも増して必要条件であるということは、提案されている文字が何を指しているのか(アイデンティティー)は携帯各社のベンダー特殊文字符号と UCS 符号間のマッピングによってある程度決定されるという面があり、そのためこの提案には「ソース表」が含まれています。文字のアイデンティティーをソースとのマッピングを利用して決めるという方法は東アジア言語の統一表象文字の符号化に使われた方法を踏襲しています。

ここで注記すべきことは、コアセットの全ての文字を符号化することはインターオペラビリティ条件を満たす為には必須であることです。そうしないとラウンド・トリップ変換やデータ交換の時にデータ・ロスが起こってしまうからです。

ソースの文書

Unicode コンソーシアムは emoji_sources.txt (N3585=L2/09-078) という文書を ISO/IEC 10646 の中の規定内容の一部として位置づけることを要請します。これは参照によって行うことが出来ます。

参照文献

- <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=BVLIXDYd&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>

考慮の要素

この提案を作成するに当たっては次のような要素を考慮に入れています。

1. **文字セットの完全性:** 提案されたシンボルは各携帯会社間のシンボルとの完全なラウンド・トリップ変換を念頭に置いたものであり、現存 Unicode 文字との統一（ユニフィケーション）も考慮しています。これはインターオペラビリティを確保するのに必要なことです。全ソース情報に関しては emoji_sources.txt file (N3585=L2/09-078) をご覧ください。
2. **ソース・セパレーションの原則:** もし携帯1社でも二つの文字を別のものとして扱っていれば —— それが JIS 標準文字セットを含む文字集合のどれかで行われていれば —— その二つの文字は UCS に於いて二つの別の文字として符号化されなければいけない。これは中・日・韓語統一表象文字の符号化の時の原則と似ているもので、例外なく厳守しなければならない原則です。
3. **文字の再使用:** 現存する UCS のシンボルを不適當なとき以外はできるだけ使うこと。これには「近い将来」ISO/IEC 10646 AMD6 と Unicode 5.2 に取り組むことが決定している文字との統一も含まれます。特筆すべきは幾つかの文字は ARIB 文字セット内の文字と統一すべきものであることです。
4. **一般性のある文字を区別する:** もし UCS に関連したシンボルのセットがあるとします。ですが、このセットのどの文字も絵文字のセットにある文字ほどの一般性がないとします。この場合には新しく1文字を追加することになります。例えば、絵文字の中の三日月シンボルは各社間で「満ちていく三日月」なのか「欠けていく三日月」なのか区別していません。（訳注：UCS では「満ちていく三日月」と「欠けていく三日月」を区別していますが、一般的な三日月はありません。この場合は現存する UCS の「三日月」のどちらとも統一せず、そのどちらでもない「三日月」の新文字を提案します。）
5. **色とアニメーション:** 色やアニメーションは符号化された文字の一部ではありません。文字は個々の実装に於いて関連した色やアニメーションがあることがあります。しかし、それはその文字のアイデンティティーの本質的な部分ではありません。
6. **現存するクロス・マッピング表:** この提案は上記のマッピング表に出来る限り準拠しています。例えば、携帯各社の12星座のシンボルはそのイメージが大分違っていますが、統一した文字として扱っています。何故かという、これらの文字は各キャリアで一つのまとまったセットに属するもので、キャリア間でクロス・マッピングが行われているからです。又、イメージが非常に違って、意味も関連付けの出来ない少数の文字は別文字として扱われています。例えば地下鉄の絵文字の一つ 'M' シンボルは地下鉄の電車の絵文字から区別されています。（このような場合）キャリア間の Shift_JIS 上のラウンド・トリップ・マッピングは Unicode と各キャリアの Shift_JIS のマッピング表で適当なフォールバックを使えば維持できることになります。

シンボルの識別名

この提案を開発中に不変の内部用の識別名、例えば ALARM CLOCK シンボルに e-02A のような識別名を割り振りました。これらの識別名はあくまで開発途上で参考に使用されたものです。

コードポイントの割り振り

現存の文字と緊密に関連している幾つかのシンボルは空きのスペースがあれば、同じ文字ブロックや関連した「補助」ブロックに割り当てられます。殆どのシンボルは SMP (補助多言語面) の新しいブロックに割り当てるように提案します。絵文字用に BMP (基本多言語面) に新規のブロックを作ることはしていません。674 のシンボルの内、9 文字は BMP で符号化し、残りの 665 文字は SMP で符号化するように提案します。

特別なまれにしか使用されないキャリア独自のシンボルはこの提案では 絵文字コンパティビリティ・シンボルのブロックに符号化するように提案しています。これらの絵文字はインターオペラビリティに向けて文字セットを完全ににする為に必要なものですが、特定のグリフや名前でなく、ソースファイルに於けるマッピングによって認定されるようになっています。

Unicode プロパティ

殆どのシンボルは標準的な Unicode プロパティを伴って提案されており、Bidi_Mirroring=False となっています。例えば、次のようなものです。

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

例外:

- 一つのシンボル (e-B08 LOOPED LENGTH MARK) は 句読点文字 (gc=Pd) として提案されています。(この文字は gc=Pd の素性を持つ U+3030 WAVY DASH と関連しています。)
- 幾つかのシンボルはコンパティビリティ分解されているものがあります。これらはチャート (N3583) に注記されてあります。この分解された文字のセットでは Unicode/ISO 10646 において既に確立された方法を使っています。

文字の照合順序

デフォルト照合順序 (DUCET/ISO 14651) は Dingbats の例に倣います。例外は

- e-B08 LOOPED LENGTH MARK は U+3030 WAVY DASH と同じにソートされること
- 分解された囲み文字はいつもの様に正規化された同等文字とは3次の相違を伴ってソートされます。

提案要約文書

[訳注: 下記の文書は ISO/IEC JTC 1/SC2/WG2 に提案した正式な願書です。これは翻訳の対象外とします。]

**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 | <input type="checkbox"/> | B.1-Specialized (small collection) | <input type="checkbox"/> | B.2-Specialized (large collection) | <input checked="" type="checkbox"/> |
| C-Major extinct | <input type="checkbox"/> | D-Attested extinct | <input type="checkbox"/> | E-Minor extinct | <input type="checkbox"/> |
| F-Archaic Hieroglyphic or Ideographic | <input type="checkbox"/> | G-Obscure or questionable usage symbols | | <input type="checkbox"/> | |
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 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? **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:

[Redacted]

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:

[Redacted]

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

[Redacted]

If Yes,
reference:

[Redacted]

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)

[Redacted]

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:

[Redacted]