Network Working Group
Request For Comments: 1815
Category: Informational

M. Ohta
Tokyo Institute of Technology

July 1995

Character Sets ISO-10646 and ISO-10646-J-1

Status of this Memo

This memo provides information for the Internet community. This memo does not specify an Internet standard of any kind. Distribution of this memo is unlimited.

Abstract

Though the ISO character set standard of ISO 10646 is specified reasonably well about European characters, it is not so useful in an fully internationalized environment.

For the practical use of ISO 10646, a lot of external profiling such as restriction of characters, restriction of combination of characters and addition of language information is necessary.

This memo provides information on such profiling, along with charset names to each profiled instance.

Though all the effort is done to make the resulting charset as useful 10646 based charset as possible, the result is not so good. So, the charsets defined in this memo are only for reference purpose and its use for practical purpose is strongly discouraged.

Introduction

This memo describes two text encoding schemes based on ISO 10646 [10646].

As ISO 10646 specifies too little about how text is visualized, to practically use ISO 10646, it is necessary to restrict the standard minimally and then add some amount of profiling information.

For ISO 2022 [ISO2022] based national standards, sufficient profiling information is provided by national standardization bodies, but, for ISO 10646, such a profiling is not yet provided.

As the profiling of ISO 10646 largely affects which character or combination of characters could be properly displayed, changes of profiling of ISO 10646 are as significant as additions of new character sets of ISO 2022.

That is, it's impractical to support the entirety of ISO 10646 (new restriction or profiling can always be added), so a client needs to know whether some restriction or profiling is being used before it can decide whether to display the body part. Thus, it is necessary to provide multiple charset names to each variation of ISO 10646.

For example, in Japan with Japanese windows NT, only those Han characters already supported by MS Kanji code (mostly equivalent to JIS X 0208 [JISX0208]) can be displayed, because no other font pattern is commonly provided.

The other problem of ISO 10646 for Han characters is that, to display them in quality required for daily plain text processing in China/Japan/Korea, it is necessary to add profiling information on which one of Chinese/Japanese/Korean the text is using. It should be noted that this feature makes multilingual mixed Chinese/Japanese/Korean text with ISO 10646 impractical.

Also, just as [RFC1521] was unclear about how bi-directionality should be supported with "ISO-8859-6" and "ISO-8859-8" which was corrected by [RFC1556], it is also unclear how bi-directionality could be supported with ISO 10646. There are too much ways to support bi- directionality. So, until some bi-directionality mechanism(s) becomes widely supported, it is necessary to exclude characters for languages which requires bi-directionality support from the minimal variation. It should be noted that, though ISO 10646 is intended to be free from long term states, save for some profiling information, introduction of bi-directionality with ISO 10646 do requires the long term states.

Combining characters also cause problems. In many countries where combining characters based on [ISO2022] is used, there are restrictions on how combining characters are ordered [TIS]. Without such restriction, the result of combination is completely meaningless which is the current state of ISO 10646. That is, if some combination is allowed in some implementation while the other does not support it, communication between them is difficult unless ISO 10646 is profiled to be least common set of widely supported combinations. So, again, until combination restriction will be developed for each language, it is necessary to exclude characters for such languages from the minimal variation.

Conjoining characters also, may or may not be supported, which requires another profiling.

According to those considerations, this memo defines two variations of ISO 10646. They are "ISO-10646" as the minimal basic variation and "ISO-10646-J-1" as the variation which could be useful in Japan.

Finally, this memo, by no means, promotes the use of ISO 10646 on the Internet. It's use is strongly discouraged, when there are other charsets which can encode the same information, Families of ISO 10646 based charsets, like ISO 2022 based charsets, only forms set of mutually incompatible encoding systems and, unlike ISO 2022 based charsets [2022INT], they can not be merged together to be the single world wide charset.

Description of "ISO-10646"

ISO-10646 is profiled to be the most basic part of the family of encodings based on ISO 10646 and contains the following minimal graphic characters:

collection number and name	positions	further restriction
1 BASIC LATIN	0020-007E	
2 LATIN-1 SUPPLEMENT	00A0-00FF	

C0 and C1 control characters may also be used as specified in the section 16 of ISO 10646.

The text with "ISO-10646" encodes text in 16 bit big endian form.

As no combining characters are included, "ISO-10646" can be used with applications at implementation level 1.

Left-to-right directionality should be used.

The encoding is implemented by Windows/NT.

For practical communication, use of "ISO-10646" is discouraged. "ISO-8859-1" [RFC1345] should be used instead.

Description of "ISO-10646-J-1"

ISO-10646-J-1 is profiled to be useful for Japanese PC users who use Japanese version of Windows/NT and contains the following graphic characters:

collection number and name	positions	further restrictions
1 BASIC LATIN	0020-007E	
2 LATIN-1 SUPPLEMENT	00A0-00FF	
8 BASIC GREEK	0370-03CF	
10 CYRILLIC	0400-04FF	
32 GENERAL PUNCTUATION	2000-206F	See note 1, below.
39 MATHEMATICAL OPERATORS	2200-22FF	See note 1, below.
44 BOX DRAWING	2500-257F	
49 CJK SYMBOLS AND PUNCTUATION	3000-303F	See note 1, below.
50 HIRAGANA	3040-309F	
51 KATAKANA	30A0-30FF	
60 CJK UNIFIED IDEOGRAPHS	4E00-9FFF	See note 1, below.
62 CJK COMPATIBILITY IDEOGRAPHS	F900-FAFF	See note 1, below.
66 CJK COMPATIBILITY FORMS	FE30-FE4F	
69 HALFWIDTH AND FULLWIDTH FORMS	FF00-FFEF	

Note 1: Most of the characters are excluded. That is, only those characters of JIS X 0208 [JISX0208] are included. The reason is that the Japanese version of Windows/NT have fonts for them only and most of the users can not read messages which contains other characters.

C0 and C1 control characters may also be used as specified in the section 16 of ISO 10646.

The text with "ISO-10646-J-1" encodes text in 16 bit big endian form.

Shapes of Han characters should be of Japanese Han, that is, those of column "J" in section 26 of ISO 10646.

As no combining characters are included, "ISO-10646-J-1" can be used with applications at implementation level 1.

Characters in "HALFWIDTH AND FULLWIDTH FORMS" compared to be different characters to the normal width characters.

When text is displayed horizontally, left-to-right directionality should be used.

For practical communication, use of "ISO-10646-J-1" is discouraged. ISO-2022-JP" [2022JP] should be used instead.

MIME Considerations

The names given to the character encoding methods described in this memo are, respectively, "ISO-10646" and "ISO-10646-J-1". This name is intended to be used in MIME messages as follows:

Content-Type: text/plain; charset=iso-10646

The ISO-10646 and ISO-10646-J-1 encoding are in 16-bit form, so it is often necessary to use a Content-Transfer-Encoding header. Base64 should be useful.

The ISO-10646 and ISO-10646-J-1 may also be used in MIME Part 2 headers [RFC1522]. The "B" encoding should be used with them.

References

[10646] International Organization for Standardization (ISO), "Universal Multiple-Octet Coded Character Set (UCS)", International Standard, Ref. No. ISO/IEC 10646-1:1993 (E).

[2022INT] (An Internet Draft "draft-ohta-text-encoding-*.txt" may be available).

[2022JP] Murai, J., Crispin, M., and E. van der Poel, "Japanese Character Encoding for Internet Messages", RFC 1468, June 1993.

[ISO2022] International Organization for Standardization (ISO), "Information processing -- ISO 7-bit and 8-bit coded character sets -- Code extension techniques", International Standard, Ref. No. ISO 2022-1986 (E).

[JISX0208] Japanese Standards Association, "Code of the Japanese graphic character set for information interchange", JIS X 0208-1990.

[RFC1345] Simonsen, K., "Character Mnemonics & Character Sets", RFC-1345, Rationel Almen Planlaegning, June 1992.

[RFC1521] Borenstein, N., and Freed, N., "MIME (Multipurpose Internet Mail Extensions) Part One: Mechanisms for Specifying and Describing the Format of Internet Message Bodies", RFC 1521, September 1993.

[RFC1522] Moore, K., "MIME (Multipurpose Internet Mail Extensions) Part Two: Message Header Extensions for Non-ASCII Text", RFC 1522, September 1993.

[RFC1556] Nussbacher, H., "Handling of Bi-directional Texts in MIME" RFC 1556, Israeli Inter-University Computer Center, December 1993.

[TIS] Thai Industrial Standard for Thai Character Code for Computer, TIS 620-2533:1990.

Security Considerations

Security issues are not discussed in this memo.

Author's Address

Masataka Ohta
Tokyo Institute of Technology
2-12-1, O-okayama, Meguro-ku,
Tokyo 152, JAPAN

Phone: +81-3-5499-7084
Fax: +81-3-3729-1940
EMail: mohta@cc.titech.ac.jp