Chinese Character Encoding for Internet Messages

Status of this Memo

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

Abstract

This memo describes methods of transporting Chinese characters in Internet services which transport text, such as electronic mail [RFC-822], network news [RFC-1036], telnet [RFC-854] and the World Wide Web [RFC-1866].

Introduction

As the use of Internet covers more and more Chinese people in the world, the need has increased for the ability to send documents containing Chinese characters on the Internet. The methods described in this document provide means of transporting existing Chinese character sets as well as leaving space for future extension.

This document describes two encodings, ISO-2022-CN and ISO-2022-CN-EXT. These are designed with interoperability in mind and are encouraged in this document for current Chinese interchange; they are 7-bit, support both simplified and traditional characters using both GB and CNS/Big5, and do not impose any unusual quoting requirements on ASCII characters.

As important related issues, this document gives detailed descriptions of the two encodings CN-GB and CN-Big5, and a brief description of ISO/IEC 10646 [ISO-10646]. CN-GB and CN-Big5 are
ISO-10646 is the universal multi-octet character set defined by ISO; we feel that in the future it may become the preferred technology for Chinese documents and electronic mail when it is widely available.

Specification


1.1. Description


ISO-2022-CN-EXT is a superset of ISO-2022-CN that additionally supports other GB character sets and planes of CNS 11643.

Since ISO-2022-CN and ISO-2022-CN-EXT are 7-bit encodings, they do not require the 8-bit SMTP extensions. ISO-2022-CN supports all the Chinese characters that appear in Big5 [BIG5].

1.2. ISO-2022-CN

The starting code of ISO-2022-CN is ASCII. ASCII and Chinese characters are distinguished by designations (ESC sequences) and shift functions.

Designations define the Chinese character sets used in the text. There are three kinds of designations: SOdesignation, SS2designation and SS3designation.

The SOdesignation is in the form ESC $ ) <F>, where <F> is the "final character" assigned to the character set by ISO (refer to the ISO registry [ISOREG] for more details). The SS2designation is in the form ESC $ * <F>, and the SS3designation is in the form ESC $ + <F>. A designation overrides any previous designation for subsequent bytes in the text.

There are four kinds of shifts: SI, SO, SS2 and SS3. Shift functions specify how to interpret the subsequent bytes.

The shift SI (one byte with hexadecimal value 0F) declares that subsequent bytes are interpreted in ASCII.
The shift SO (one byte with hexadecimal value 0E) declares that subsequent bytes are interpreted in the character set defined by SOdesignation.

The shift SS2 (two bytes with hexadecimal values 1B 4E) declares that the subsequent TWO bytes are interpreted in the character set defined by SS2designation, after which the previous interpretation (from SI or SO) is restored.

The shift SS3 (two bytes with hexadecimal values 1B 4F) declares that the subsequent TWO bytes are interpreted in the character set defined by SS3designation, after which the previous interpretation (from SI or SO) is restored.

The escape sequences, shift functions and character sets used in an ISO-2022-CN text are as follows:

<table>
<thead>
<tr>
<th>Character sets</th>
<th>Shift in with</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASCII</td>
<td>SI</td>
</tr>
<tr>
<td>GB 2312, CNS 11643-plane-1</td>
<td>SO</td>
</tr>
<tr>
<td>CNS 11643-plane-2</td>
<td>SS2</td>
</tr>
</tbody>
</table>

ESC $ ) A Indicates the bytes following SO are Chinese characters as defined in GB 2312-80, until another SOdesignation appears

ESC $ ) G Indicates the bytes following SO are as defined in CNS 11643-plane-1, until another SOdesignation appears

ESC $ * H Indicates the two bytes immediately following SS2 is a Chinese character as defined in CNS 11643-plane-2, until another SS2designation appears

If there are any GB or CNS characters on a line, a designation for the corresponding character set must be used so that each line has its own character set information and the text can be displayed correctly when scroll back in a window. Also, there must be a shift to ASCII (SI) before the end of the line (i.e., before the CRLF). In other words, each line starts in ASCII, and ends in ASCII.

Example: the hex sequence

1b 24 29 41 0e 3d 3b 3b 3b 3b 1b 24 29 47 47 28 5f 50 0f

represents the Chinese word for "Interchange" (jiao huan) twice;
the first time in simplified form using GB-2312 (the 3d 3b 3b 3b sequence above), and the second time in traditional form using CNS-11643 (the 47 28 5f 50 sequence above). The sequence 1b 24 29 41 is the SO designation for GB-2312, the 0e is SO to switch to Chinese from ASCII, the 1b 24 29 47 is the SO designation for CNS-11643 plane 1, and finally the 0f is the SI to return to ASCII at the end of the line.

The name given to this character encoding is "ISO-2022-CN". This name is intended to be used as the "charset" parameter in MIME [MIME-1, MIME-2] messages.

Content-Type: text/plain; charset=iso-2022-cn

The ISO-2022-CN encoding is already in 7-bit form, so it is not necessary to use a Content-Transfer-Encoding header.

Other restrictions are given in the "Formal Syntax of ISO-2022-CN" (Section 7.1 of this document).

1.3. ISO-2022-CN-EXT

ISO-2022-CN-EXT supports all characters in existing GB, Big5 and CNS 11643 character sets.

The escape sequences, shift functions and character sets used in an ISO-2022-CN-EXT text are as follows:

<table>
<thead>
<tr>
<th>Character sets</th>
<th>Shift in with</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASCII</td>
<td>SI</td>
</tr>
<tr>
<td>GB 2312, GB 12345, CNS 11643-plane-1, ISO-IR-165</td>
<td>SO</td>
</tr>
<tr>
<td>GB 7589, GB 13131, CNS 11643-plane-2</td>
<td>SS2</td>
</tr>
<tr>
<td>GB 7590, GB 13132 or other new GBs, CNS 11643-plane-3 or higher planes of CNS 11643</td>
<td>SS3</td>
</tr>
</tbody>
</table>

Note: Currently, there are some GB sets that have not been registered in ISO. Here <X7589>, <X7590>, <X12345>, <X13131> and <X13132> represent the final character that will be assigned by ISO for those sets. These GB sets shall only be used once these final characters are assigned.
ESC $) A Indicates the bytes following SO are Chinese characters as defined in GB 2312-80, until another SO designation appears.

ESC $ * <X7589> Indicates the two bytes immediately following SS2 is a Chinese character as defined in GB 7589-87 [GB-7589], until another SS2 designation appears.

ESC $ + <X7590> Indicates the two bytes immediately following SS3 is a Chinese character as defined in GB 7590-87 [GB-7590], until another SS3 designation appears.

ESC $) <X12345> Indicates the bytes following SO are as defined in GB 12345-90 [GB-12345], until another SO designation appears.

ESC $ * <X13131> Indicates the two bytes immediately following SS2 is a Chinese character as defined in GB 13131-91 [GB-13131], until another SS2 designation appears.

ESC $ + <X13132> Indicates the two bytes immediately following SS3 is a Chinese character as defined in GB 13132-91 [GB-13131], until another SS3 designation appears.

ESC $) E Indicates the bytes following SO are as defined in ISO-IR-165 (for details, see section 2.1), until another SO designation appears.

ESC $) G Indicates the bytes following SO are as defined in CNS 11643-plane-1, until another SO designation appears.

ESC $ * H Indicates the two bytes immediately following SS2 is a Chinese character as defined in CNS 11643-plane-2, until another SS2 designation appears.

ESC $ + I Indicates the immediate two bytes following SS3 is a Chinese character as defined in CNS 11643-plane-3, until another SS3 designation appears.
Indicates the immediate two bytes following SS3 is a Chinese character as defined in CNS 11643-plane-4, until another SS3 designation appears.

Indicates the immediate two bytes following SS3 is a Chinese character as defined in CNS 11643-plane-5, until another SS3 designation appears.

Indicates the immediate two bytes following SS3 is a Chinese character as defined in CNS 11643-plane-6, until another SS3 designation appears.

Indicates the immediate two bytes following SS3 is a Chinese character as defined in CNS 11643-plane-7, until another SS3 designation appears.

As in ISO-2022-CN, each line starts in ASCII, and ends in ASCII, and has its own designation information before any Chinese characters appear.

The name given to this character encoding is "ISO-2022-CN-EXT". This name is intended to be used as the "charset" parameter in MIME messages.

Content-Type: text/plain; charset=ISO-2022-CN-EXT

The ISO-2022-CN-EXT encoding is also in 7-bit form, so it is not necessary to use a Content-Transfer-Encoding header.

Other restrictions are given in the "Formal Syntax of ISO-2022-CN-EXT" (Section 7.2 of this document).

1.4. How to Support Big5 or other internal codesets with ISO-2022-CN and ISO-2022-CN-EXT

Since there are many different Chinese internal coding systems [CJKINF], such as EUC GB, Big5, CCCII (an encoding for library systems mainly used in Taiwan), GBK (the new standard specification for Chinese internal code, also is the codepage for Microsoft simplified Chinese Windows 95) etc., ISO-2022-CN and ISO-2022-CN-EXT, which are 7-bit and will not lose information during communication among different codesets, facilitate interchange between the various Chinese coding systems in the Internet.
For instance, ISO-2022-CN and ISO-2022-CN-EXT can be used to support the popular Big5 codeset, because the first two planes of CNS-11643 contain the same Chinese characters as Big5’s "common part" except two duplicate characters. By the "common part" we mean the part that is not specific to any Big5 vendor, consisting of 5401 more frequently used characters in Big5 range 0xA440-0xC67E, 7652 less frequently used characters in Big5 range 0xC940-0xF9D5, and 441 other symbols in Big5 range 0xA140-0xA3E0, as defined in Institute for Information Industry’s (III) technical report C-26 (see also [Big5]). The appendix of this document presents a conversion table for converting Big5 into CNS-11643, including specific extensions of some popular vendors. For other extensions, vendors and implementors of Big5 products are ENCOURAGED to create detailed conversion tables, in order to increase interoperability between different coding systems.

Public domain software (binary or C source code) for conversion between Big5 and CNS-11643 is available on many Internet sites. At the time of this writing, the following FTP sites and software are advertised:

1) Beijing:
   (IP address: 166.111.1.6)

2) Xi’an:
   ftp://ftp.xanet.edu.cn
   /pub/chinese-soft/unix/convert/BeTTY-1.534.tar.gz
   (IP address: 202.112.11.131)

3) Taiwan:
   ftp://ftp.seed.net.tw/Pub/Chinese/DOS/code-convert/chcode.zip
   (IP address: 140.92.1.65)

4) US:
   ftp://ftp.ifcss.org/pub/software/unix/convert/BeTTY-1.534.tar.gz
   (IP address: 128.123.1.55)

5) Japan:
   (IP address: 192.31.197.99)
2. 8-bit Chinese encodings: CN-GB and CN-Big5

The CN-GB and CN-Big5 MIME charsets are defined below.

Note: the use of 8-bit character sets requires the use of either an 8-to-7 Content-Transfer-Encoding mechanism such as "BASE64" or "QUOTED-PRINTABLE" if the network is not 8-bit clean, or the 8-bit SMTP extensions [SMTPEXT] with the "8BIT" Content-Transfer-Encoding on 8-bit clean networks. Otherwise, an 8-bit message that passes through a 7-bit mailer is likely to have the 8th bit truncated, resulting in an unreadable message. Although "just send 8-bit data" has been common practice in the past, it is incorrect according to the Internet standards and causes interoperability problems.

2.1. CN-GB

E-mail using CN-GB characters is sent in this way:

GB 2312-80 characters are used with ASCII characters, not GB 1988-89 [GB-1988].

GB 2312-80 is also 7-bit, to avoid conflicting with ASCII. If the character is from GB 2312-80, the MSB (bit-8) of each byte is set to 1, and therefore becomes a 8-bit character. Otherwise, the byte is interpreted as ASCII. This constructs a character set named "GB Internal Code".

This method is also adopted in the .gb files in the Internet.

To use this character scheme with MIME, CN-GB is used as the value for the charset parameter:

Content-Type: text/plain; charset=cn-gb; charset-edition=1980

Note: The "charset-edition" is a new MIME parameter described in section 4.1 of the "Specification" part of this document.

GB 12345-90 is the traditional form of GB 2312, the charset name given to this set is CN-GB-12345 with the charset-edition of 1990.

There are also character sets that can only be used with other GB sets. For example, GB 8565-88 [GB-8565] is used with GB 2312 and some other characters to form the ISO-IR-165 set (also known as GB 2312 + GB 8565.2). ISO-IR-165 contains all characters from GB 2312-80 as revised by GB 6345.1-86 and GB 8565.2-88. Its MIME charset name is CN-GB-ISOIR165 with the charset-edition of 1992.
CN-GB-12345 and CN-GB-ISOIR165 support ASCII in a similar manner to CN-GB; the MSB of Chinese characters is set to 1 to distinguish from ASCII.

Note: There are some supplementary character sets in GB, i.e. GB 7589-87, GB 7590-87, GB 13131-91 and GB 13132-91. Normally, they won't be used independently without using GB-2312 or GB-12345, so they are not necessarily to be registered. Characters in these standards could be supported with ISO-2022-CN and ISO-2022-CN-EXT. If, in the future, they need to be used with "charset" names, it is the responsibility of any interested third party (the standardization organization or anybody else) to write the necessary documents and register the charset with the IANA. It is encouraged that the charset names take the form of CN-GB-<number>, such as CN-GB-12345, where <number> is the GB standard number. A charset-edition should also be given. All CN-GB-<number> sets should be coded in 8-bit in a similar fashion to CN-GB.

To ensure interoperability, the CN-GB charset should be used whenever possible instead of a CN-GB-<number> charset.

2.2. CN-Big5

Big5 is a two-byte character set of traditional Chinese characters, widely used in Taiwan and overseas. E-mail of CN-Big5 is sent in this way:

Big5 is used with ASCII. The MSB of ASCII characters is always 0. The MSB of the first byte of a Big5 character is always 1; this distinguishes it from an ASCII character. The second byte has 8 significant bits. Therefore, CN-Big5 is an 8-bit encoding with a 15-bit codespace.

To use this character scheme with MIME, CN-Big5 is used as the value for the charset parameter:

Content-Type: text/plain; charset=cn-big5; charset-edition=1984

Note: The "charset-edition" is a new MIME parameter described in section 4.1 of the "Specification" part of this document.


ISO/IEC 10646 defines a 32bit character space with the intent to encode all characters in the world. Currently, only the lowest 16bit plane of ISO 10646, the Basic Multilingual Plane (BMP), is defined. The BMP is code-by-code identical to Unicode [Unicode 1.1]. It contains a large repertoire of Chinese characters (it currently
includes all the characters of GB 2312-80, GB 12345-90, GB 8565-89, CNS 11643’s plane 1 and 2, and part of some other standards) and therefore can be used to transport Chinese characters in the Internet community. This document does not give any details on how to do this, as this has been done elsewhere. For details of using Unicode with MIME, refer to RFC 1641 [RFC-1641], RFC 1642 [RFC-1642]. For assigned names for 10646 set, refer to STD 2--"Assigned Numbers", which is RFC 1700 [RFC-1700] currently. For more up-to-date assigned numbers, please check:

ftp://ftp.isi.edu/in-notes/iana/assignments/character-sets

4. Two New MIME parameters

Here we define two new MIME parameters to be used with "charset" parameters.

4.1. "charset-edition"

This parameter is used after the MIME "charset" parameter, using four digits (AD) to indicate what the year of edition is for the character set standard shown in "charset". Its use is optional. Implementations should ignore this parameter unless the implementation has specific support for that particular character set edition.

The reason for defining this parameter is that there are often differences in the defined characters between editions of a character set standard. Sometimes, the difference can not be ignored, otherwise implementations would have problems when processing it. There are only two ways to indicate this difference, in the current MIME syntax. One way is to indicate the edition in the charset name, such as CN-GB-1988-80 (the 1980’s edition of GB 1988). The other way is to define a new optional parameter such as "charset-edition". The latter way is better because receiving applications that can only process an older edition can still recognize the character set and offer to display the text in the older edition. This display may have a few mistakes, but it is better than refusing to display any text at all or defaulting to an inappropriate character set such as US-ASCII or ISO-8859-1.

4.2. "charset-extension"

This parameter is also used after the MIME "charset" parameter. It is case-insensitive and optional, and any value of this parameter should be registered in IANA. Unregistered value should start with "x-" as with any MIME extension-token. Implementations should ignore this parameter unless the implementation has specific support for
that particular character set extension.

A character set extension has displayed glyphs for code points that are not assigned in the character set, for example, vendor-specific extensions of standard character sets. This parameter provides the option of using these extensions. Although character set extensions may cause interoperability problems, we recognize the existence of such extensions.

For example:

Content-Type: text/plain; charset=CN-Big5; charset-edition=1984; charset-extension=ETen-2.00.03-DOS

This may indicate Eten company’s extension of Big5: ETen 2.00.03 for DOS, assuming that "ETen-2.00.03-DOS" is registered with the IANA.

4.3. Formal Syntax:

The following changes and additions are made to the MIME syntax:

```plaintext
charset-edition   := "charset-edition" "=" 4DIGIT
                    ; year of edition in four digits

charset-extension := "charset-extension" "=" extension-token
```

5. Background Information

5.1. Writing systems and their encodings in Chinese-speaking nations and regions

The mainland provinces of China use simplified Chinese character in daily life. GB is the standard electronic character set. It is the main means for communications between people who share simplified Chinese characters in the world.

Taiwan uses traditional Chinese characters in daily life. CNS-11643 is the formal character set for information interchange in Taiwan; however, Big5, a widely-used character set of traditional Chinese characters, is the de-facto internal code standard in Taiwan.

Hong Kong uses traditional Chinese characters in daily life, but uses both GB and Big5 in electronic form, because Hong Kong people often communicate with people in all of China’s provinces.

Singapore seldom uses Chinese characters, and uses the simplified form when Chinese characters are used. In electronic form, Unicode is more popular, however GB is also used.
5.2. Miscellaneous information about Chinese character sets

The GB 1988-89 character set is identical to ISO 646 [ISO-646] except for currency symbol and tilde. The currency symbol and the tilde are replaced by the Yuan sign and the overline. This set is GB’s variant of ISO 646. This character set and CNS 5205 [CNS-5205] are not encouraged for use in the Internet, since ASCII combined with GB 2312 or CNS 11643-plane 1 and plane 2 contains all the characters in them.

The GB 2312-80 character set consists of simplified Chinese characters, digits, and the Latin, Greek and Russian alphabets, and some other symbols; in all, 7445 characters. Each character is represented with two bytes.

GB 13000-95 [GB-13000] is GB’s variant of ISO 10646. However, for interoperability in the Internet, assigned names for ISO 10646 are encouraged instead.

Currently both sides of the Taiwan Straits are cooperating closely in promoting the use of ISO 10646’s BMP and in continuing its development together with other organizations under ISO.

5.3. Miscellaneous implementation information

For maximum interoperability, implementations SHOULD at least support sending and receiving ISO-2022-CN. Supporting all registered character sets in ISO-2022-CN-EXT is greatly encouraged.

To meet the current usage, support of CN-GB (the status quo for simplified Chinese e-mail) or CN-Big5 (the status quo for traditional Chinese e-mail) may be necessary. However, it is not reliable to send documents directly with these internal codes, therefore sending ISO-2022-CN message is always encouraged whenever possible.

To the maximum extent possible, implementations should be capable of receiving messages in any of the encodings described in this document, even if they only transmit messages in one form.

Preferably the implementation should display the characters with glyphs appropriate to the typographic tradition that is implied in the encoding of the received text. Implementation may also translate these encodings to the encoding that its platform supports.

The human user (not implementor) should try to keep lines within 80 display columns, or, preferably, within 75 (or so) columns, to allow insertion of ">" at the beginning of each line in excerpts. Each Chinese character takes up two columns, and the shift sequences do
not take up any columns. The implementor is reminded that Chinese characters take up two bytes and should not be split in the middle to break lines for displaying, etc.

Freely available fonts of Chinese characters:

Beijing:

Xi’an:

Taiwan:

Hong Kong:

Singapore:

US:
   http://ccic.ifcss.org/www/pub/software/fonts/

6. X.400 Considerations

X.400 has the ability of carrying different character sets in a message by using the body part "GeneralText" defined by ISO/IEC-10021-7 [ISO-10021].

The X.400 ASN.1 definition of the GeneralText body part is:

   general-text-body-part EXTENDED-BODY-PART-TYPE
   PARAMETERS GeneralTextParameters IDENTIFIED BY id-ep-general-text
   DATA GeneralTextData ::= id-et-general-text

GeneralTextParameters ::= SET OF CharacterSetRegistration

CharacterSetRegistration ::= INTEGER (1..32767)

GeneralTextData ::= GeneralString

Therefore, to use ISO-2022-CN, set the "CharacterSetRegistration" part as { 6 58 171 172 }, and add an ESC sequence of ESC (B (three bytes, hexadecimal values: 1B 28 42) before the beginning of each
Similarly, to use ISO-2022-CN-EXT, set the registered numbers of all character sets in the "CharacterSetRegistration" part and add ESC { B at the beginning of each line. For the registered numbers, please refer to ISO registry. In addition to the character sets supported by ISO-2022-CN, currently registered numbers are:

<table>
<thead>
<tr>
<th>Character Set</th>
<th>Registered Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISO IR 165 (GB 2312+GB 8565.2)</td>
<td>165</td>
</tr>
<tr>
<td>CNS 11643-plane 3</td>
<td>183</td>
</tr>
<tr>
<td>CNS 11643-plane 4</td>
<td>184</td>
</tr>
<tr>
<td>CNS 11643-plane 5</td>
<td>185</td>
</tr>
<tr>
<td>CNS 11643-plane 6</td>
<td>186</td>
</tr>
<tr>
<td>CNS 11643-plane 7</td>
<td>187</td>
</tr>
</tbody>
</table>

176 is the registered number for the BASESET of ISO/IEC 10646-1:1993 UCS-2 with implementation level 3, Escape sequence of ESC % / E (four bytes, hexadecimal values 1B 25 2F 45) indicates starting of this codeset.

For CN-GB and CN-Big5 character sets, there are no formal methods that could be used in X.400 yet.

For detail about X.400 use of character sets, please refer to RFC 1502 [RFC-1502].

The notational conventions used here are identical to those used in RFC 822.


\[
\begin{align*}
\text{body} & ::= * ( \text{ascii_line} / \text{c_line} ) \\
\text{ascii_line} & ::= *\text{char} \text{CRLF} \\
\text{c_line} & ::= *\text{char} 1*(1*\text{designation} 1*(*\text{char} 1*\text{c_text} *\text{char})) \text{CRLF} \\
\text{designation} & ::= \text{SOdesignation} / \text{SS2designation} \\
\text{SOdesignation} & ::= \text{ESC} "$" (\text{finalchar_for_SO}) \\
\text{SS2designation} & ::= \text{ESC} "$" "$" (\text{finalchar_for_SS2}) \\
\text{finalchar_for_SO} & ::= "A" / "G" \\
\text{finalchar_for_SS2} & ::= "H" \\
\text{c_text} & ::= 1* ( \text{SO-SI-segment} / \text{SS2segment} ) \\
\text{SO-SI-segment} & ::= \text{SO} 1*\text{char} *\text{designation} *\text{c_segment} \text{SI} \\
\text{c_segment} & ::= 1* ( \text{c_char} / \text{SS2segment} ) \\
\text{SS2segment} & ::= \text{SS2} \text{c_char} \\
\text{c_char} & ::= \text{one_of_94 one_of_94} \\
\text{ESC} & ::= \text{<ISO-646 ESC, escape>} \quad ; \quad (33, 27.) \\
\text{SI} & ::= \text{<ASCII SI, shift in>} \quad ; \quad (17, 15.) \\
\text{SO} & ::= \text{<ASCII SO, shift out>} \quad ; \quad (16, 14.) \\
\text{SS2} & ::= \text{<ISO 2022 Single_shift two>} \quad ; \quad (33 116, 27 78.) \\
\text{one_of_94} & ::= \text{<any char in 94_char_set>} \quad ; \quad (41-176, 33-126.) \\
\text{char} & ::= \text{<any char in 96_char_set>} \quad ; \quad (40-177, 30-127.)
\end{align*}
\]

body ::= * ( ascii_line / c_line )

ascii_line ::= *char CRLF

c_line ::= *char 1*(1*designation 1*(char 1*c_text *char)) CRLF
designation ::= SOdesignation / SS2designation / SS3designation

SOdesignation ::= ESC "$" )" finalchar_for_SO

SS2designation ::= ESC "$" *" finalchar_for_SS2

SS3designation ::= ESC "$" +" finalchar_for_SS3

finalchar_for_SO ::= "A" / <X12345> / "G" / "E"

finalchar_for_SS2 ::= <X7589> / <X13131> / "H"

finalchar_for_SS3 ::= <X7590> / <X13132> / "I" / "J" / "K" / "L"

finalchar_for_SS3 / "M"

c_text ::= 1* ( SO-SI-segment / SS2segment / SS3segment )

SO-SI-segment ::= SO 1*c_char *designation *c_segment SI

c_segment ::= 1* ( c_char / SS2segment / SS3segment )

SS2segment ::= SS2 c_char

SS3segment ::= SS3 c_char

c_char ::= one_of_94 one_of_94

; ( Octal, Decimal.)

ESC ::= <ISO-646 ESC, escape> ; ( 33, 27.)

SI ::= <ASCII SI, shift in> ; ( 17, 15.)

SO ::= <ASCII SO, shift out> ; ( 16, 14.)

SS2 ::= <ISO 2022 Single_shift two> ; ( 33 116, 27 78.)

SS3 ::= <ISO 2022 Single_shift three>; ( 33 117, 27 79.)

one_of_94 ::= <any char in 94_char set> ; ( 41-176, 33-126.
8. Registration of New "charset"s and New MIME parameter

8.1. This document defines the following MIME "charset" names for Chinese text:

CN-GB, CN-Big5
CN-GB-12345
CN-GB-ISOIR165

8.2. This document defines two new MIME parameters:

charset-edition
charset-extension

Acknowledgments

This document is the result of cooperation in APNG-CC, the Chinese Character sub-working group of the I18N/L10N (Internationalization and Localization) working group of APNG (Asia-Pacific Networking Group), coordinator Zhu Haifeng <zhf@net.tsinghua.edu.cn>. The membership of APNG-CC consists of individuals from both sides of the Taiwan Strait, HongKong, and from Singapore and other countries. We wish to thank all members of APNG-CC.

Prof. Yao Shiquan (Deputy chair of CITS--China Information Technology Standardization Technical Committee), Ms. Lin Ning (Secretary-General of CITS), Mr. Guo Chengzhong of the Office of the Joint Conference of China Economic Information, and Prof. Zhao Jingrong, Prof. Wu Jianping, Prof. Li Xing, and Mr. You Yue (Tsinghua University) and other experts from CERNET Expert Committee, Prof. Meng Qingyu (China Computer Software & Technology Services Corporation), Prof. Cao Jinwen and Mr. Yu Jun (IBM Beijing) gave a lot of support and help in many aspects.

Special thanks for the supports towards APNG-CC from Prof. Yang Tianxing (Chair of CITS).

Prof. Ding ZyKaan from Academia Sinica of Taiwan, and Mr. C. J. Cherng and Mr. C. K. Fan of III (Institute for Information Industry), Mr. Chang JingShin from Tsinghua University in Hsinchu of Taiwan, Ms. C. C. Hsu from IBM Taiwan and Ms. Tong-Lee Anita Lin from Microsoft
Taiwan gave a lot of support and contributions in APNG-CC’s work. In particular, Ms. C. C. Hsu put much effort towards completing the Appendix of this document.

We also wish to thank the following people who contributed in many ways towards this document.

Zhang Zhoucai  Martin J. Duerst  
Zhang Ling     Kenichi Handa  
Zhu Bin        Lu Chin  
Sun Yufang     Nelson Chin  
Chen Shuyi     Mao Yonggang  
Masataka Ohta  Ken Lunde  
Luo Kim Teng   Victor Cheng  
Stephen G. Simpson  Yuan Jiang  
Liu Huifang    Harald T. Alvestrand  
Qian Hualin    Jiang Lin  
Lu Ming        Emily Hsu  
Wu Jian        Zhu Shuang  
Zheng Long     Zhang Hailin  
Yonggang Zhang Feng Hui  
Yao Jian

Security Considerations

Security issues are not discussed in this memo.

Authors’ Addresses

Zhu Haifeng (HF. Zhu)  
216 Central Main Building  
Tsinghua University  
Beijing, 100084  
China  
Tel: +86-10-2561144 ext. 3492  
Fax: +86-10-2564173  
EMail: zhf@net.tsinghua.edu.cn, zhf@net.edu.cn
Hu Daoyuan (DY. Hu)
Tsinghua Networking Center
Tsinghua University
Beijing, 100084
China

Tel: +86-10-2594016
Fax: +86-10-2564173
EMail: hdy@tsinghua.edu.cn

Wang Zhiguan (ZG. Wang)
Beijing 1101 MailBox
SubCommittee 2 (SC2)
China Information Technology Standardization Technical Committee
(CITS)
Beijing, 100007
China

Tel: +86-10-4012392
Fax: +86-10-4010601

Kao Tien-cheu (TC. Kao)
I.T. Promotion Division
Institute for Information Industry (III)
Taipei
Taiwan

Tel: +886-2-5631688
Fax: +886-2-563-4209
EMail: tckao@iidns.iii.org.tw

Chang Wen-chung (WCH. Chang)
Institute for Information Industry (III)
Taipei
Taiwan

Tel: +886-2-7327771
Fax: +886-2-7370188
EMail: chung@iidns.iii.org.tw
Appendix -- Conversion Table for ISO-2022-CN (EXT) and Big5

This is a conversion table for the Chinese characters in Big5’s common part and ISO-2022-CN/-EXT, including all the vendor-specific characters from Eten, Microsoft and IBM. For conversion source and binary programs for Big5, III provides good on-line services (ftp site listed in section 1.4), and [CJKINF] is also a good reference.

### A.1. Big5 (ETen, IBM, and Microsoft version) symbol set correspondence to CNS 11643 Plane 1:

<table>
<thead>
<tr>
<th>Big5 Code</th>
<th>CNS Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>0xA140-0xA1F5</td>
<td>0x2121-0x2256</td>
</tr>
<tr>
<td>0xA1F6</td>
<td>0x2258</td>
</tr>
<tr>
<td>0xA1F7</td>
<td>0x2257</td>
</tr>
<tr>
<td>0xA1F8-0xA2AE</td>
<td>0x2259-0x234E</td>
</tr>
<tr>
<td>0xA2AF-0xA3BF</td>
<td>0x2421-0x2570</td>
</tr>
<tr>
<td>0xA3C0-0xA3E0</td>
<td>0x4221-0x4241 (ETen and Microsoft defined as reserved area)</td>
</tr>
</tbody>
</table>

### A.2. Big5 (ETen, IBM, and Microsoft version) Level 1 correspondence to CNS 11643-1992 Plane 1:

<table>
<thead>
<tr>
<th>Big5 Code</th>
<th>CNS Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>0xA440-0xACFD</td>
<td>0x4421-0x5322</td>
</tr>
<tr>
<td>0xACFE</td>
<td>0x5753</td>
</tr>
<tr>
<td>0xAD40-0xAFCF</td>
<td>0x5323-0x5752</td>
</tr>
<tr>
<td>0xAFD0-0xBBE7</td>
<td>0x5754-0x6B4F</td>
</tr>
<tr>
<td>0xBBE8-0xBE51</td>
<td>0x6B51-0x6F5B</td>
</tr>
<tr>
<td>0xBE52</td>
<td>0x6B50</td>
</tr>
<tr>
<td>0xBE53-0xC1AA</td>
<td>0x6F5C-0x7534</td>
</tr>
<tr>
<td>0xC1AB-0xC2CA</td>
<td>0x7536-0x7736</td>
</tr>
<tr>
<td>0xC2CB</td>
<td>0x7535</td>
</tr>
<tr>
<td>0xC2CC-0xC360</td>
<td>0x7737-0x782C</td>
</tr>
<tr>
<td>0xC361-0xC3B8</td>
<td>0x782E-0x7863</td>
</tr>
<tr>
<td>0xC3B9</td>
<td>0x7865</td>
</tr>
<tr>
<td>0xC3BA</td>
<td>0x7864</td>
</tr>
<tr>
<td>0xC3BB-0xC455</td>
<td>0x7866-0x7961</td>
</tr>
<tr>
<td>0xC456</td>
<td>0x782D</td>
</tr>
<tr>
<td>0xC457-0xC67E</td>
<td>0x7962-0x7D4B</td>
</tr>
</tbody>
</table>

### A.3. Big5 (ETen, IBM, and Microsoft version) Level 2 correspondence to CNS 11643-1992 Plane 2:

<table>
<thead>
<tr>
<th>Big5 Code</th>
<th>CNS Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>0xC940-0xC949</td>
<td>0x2121-0x212A</td>
</tr>
<tr>
<td>0xC94A</td>
<td>0x4442 # duplicate of Level 1’s 0xA461</td>
</tr>
<tr>
<td>0xC94B-0xC96B</td>
<td>0x212B-0x214B</td>
</tr>
<tr>
<td>0xC96C-0xC9BD</td>
<td>0x214D-0x217C</td>
</tr>
<tr>
<td>0xC9BE</td>
<td>0x214C</td>
</tr>
<tr>
<td>0xC9BF-0xC9EC</td>
<td>0x217D-0x224C</td>
</tr>
</tbody>
</table>
A.4. Big5 (ETen and IBM Version) specific numeric symbols correspondence to CNS 11643 Plane 1: (Microsoft version defined this area as UDC - User Defined Character)
A.5. Big5 (ETen and IBM Version) specific KangXi radicals correspondence to CNS 11643 Plane 1: (Microsoft version defined as UDC - User Definable Character)

- 0x0xC6BF <-> 0x0x2723
- 0x0xC6C0 <-> 0x0x2724
- 0x0xC6C1 <-> 0x0x2726
- 0x0xC6C2 <-> 0x0x2728
- 0x0xC6C3 <-> 0x0x272D
- 0x0xC6C4 <-> 0x0x272E
- 0x0xC6C5 <-> 0x0x272F
- 0x0xC6C6 <-> 0x0x2734
- 0x0xC6C7 <-> 0x0x2737
- 0x0xC6C8 <-> 0x0x273A
- 0x0xC6C9 <-> 0x0x273C
- 0x0xC6CA <-> 0x0x2742
- 0x0xC6CB <-> 0x0x2747
- 0x0xC6CC <-> 0x0x274E
- 0x0xC6CD <-> 0x0x2753
- 0x0xC6CE <-> 0x0x2754
- 0x0xC6CF <-> 0x0x2755
- 0x0xC6D0 <-> 0x0x2759
- 0x0xC6D1 <-> 0x0x275A
- 0x0xC6D2 <-> 0x0x2761
- 0x0xC6D3 <-> 0x0x2766
- 0x0xC6D4 <-> 0x0x2829
- 0x0xC6D5 <-> 0x0x282A
- 0x0xC6D6 <-> 0x0x2863
- 0x0xC6D7 <-> 0x0x286C

A.6. Big5 (ETen and Microsoft version) specific Ideographs correspondence to CNS 11643 Plane 3: (IBM version defined as UDC)

- 0xF9D6 <-> 0x4337
- 0xF9D7 <-> 0x4F50
- 0xF9D8 <-> 0x444E
- 0xF9D9 <-> 0x504A
- 0xF9DA <-> 0x2C5D
- 0xF9DB <-> 0x3D7E
- 0xF9DC <-> 0x4B5C

A.7. Big5 (ETen version only) specific symbols correspondence to CNS 11643 Plane 4:

- 0xC879 <-> 0x2123
A.8. Other Big5 specific symbols which cannot mapping to CNS 11643:

- 0xC6D8-0xC878 <-> none (ETen and IBM Version)
- 0xC87A <-> none (ETen version only)
- 0xC87C <-> none (ETen version only)
- 0xC87E-0xC8A1 <-> none (ETen version only)
- 0xC8A3-0xC8CC <-> none (ETen version only)
- 0xC8CD-0xC8D3 <-> none (ETen and IBM version)
- 0xF9DD-0xF9FE <-> none (ETen and Microsoft version)

Note: However, most of them can be mapped to GB-2312 too. For example, Big5(ETen and IBM version) Hiragana, Katakana, and Cyrillic symbols correspondence to GB-2312:

- 0xC6E7-0xC77A <-> 0x2421-0x2473 # Japanese Hiragana
- 0xC77B-0xC7F2 <-> 0x2521-0x2576 # Japanese Katakana
- 0xC7F3-0xC854 <-> 0xA7A1-0xA7C1 # Cyrillic uppercase
- 0xC855-0xC875 <-> 0xA7D1-0xA7F1 # Cyrillic lowercase

Please notice that there are also many symbols that could be supported by GB-2312, for detail, please refer to the ftp sites in section 1.4 of the "Specification" part of this document.
References


[CNS-5205] "Information processing: 7-Bit Coded Character Set For Information Interchange", CNS-5205.


[GB-2312] "Coding of Chinese Ideogram Set for Information Interchange Basic Set", GB 2312-80.

[GB-7589] "Code of Chinese Ideograms Set for Information Interchange, the 2nd Supplementary Set", UDC 681.3.048, GB 7589-87.

[GB-7590] "Code of Chinese Ideogram Set for Information Interchange, the 4th Supplementary Set", UDC 681.3.048, GB 7590-87.


[GB-13132] "Code of Chinese Ideogram Set for Information Interchange, the 5th Supplementary Set", GB 13132-91.
[ISO-646] International Organization for Standardization (ISO),  

[ISO-2022] International Organization for Standardization (ISO),  


[ISOREG] International Organization for Standardization (ISO),  
"International Register of Coded Character Sets To Be Used With Escape Sequences".


[Unicode 1.1] "The Unicode Standard, Version 1.1", Addison-Wesley, Reading, MA (to be published; the contents of this standard is currently available by combining [Unicode92], [Unicode93], and [Unicode4]).

