L2/06-284

ISO/IEC JTC 1/SC 2/WG 2 N3091

2006-04-20

ISO/IEC JTC 1/SC 2/WG 2
Universal Multiple-Octet Coded Character Set (UCS)
Secretariat: ANSI

Request For Collection Identifiers For Japanese Subsets of ISO/IEC 10646

DATE:	**2006-04-20**
SOURCE:	Japan
STATUS:	Japan National Body Request
ATTACHMENTS:	N3091-A

This document requests five new collection identifiers to be used for specifying Japanese subsets of ISO/IEC 10646 according to 12.2 of the standard.

1. Background

As of now there are basically four Japanese character set standards. They are JIS X 0201, JIS X 0208, JIS X 0212 and JIS X 0213, each of which is designed to be used in accordance with ISO/IEC 2022. On the other hand, because JIS X 0221, Japanese translation of ISO/IEC 10646, contains all these national standards as source, they can also be seen as subsets of JIS X 0221.

The most widely adopted character set in Japanese IT industry is JIS X 0208 with legacy encodings such as Shift-JIS, but it's been recognized that JIS X 0208 character set does not have a good enough coverage even for use of encoding contemporary Japanese. JIS X 0212 and JIS X 0213 were developed to satisfy such needs for good enough coverage. Because of the recent revision of JIS X 0213 in 2004 to incorporate the latest character shape standard established by National Language Council in 2000 and the legal requirement of the character set for naming newborn babies (effective 2004), commercial vendors started referring to JIS X 0213:2004 (or JIS 2004 for short). Actually in many cases it's referring only to the corresponding sub-repertoire of ISO/IEC 10646, however, there has been confusion among people whether it's referring to the accompanied legacy encoding schemes too.

Getting rid of the situation, Japan National Body received a request to add several collections to UCS, so that users and vendors in Japan can unambiguously refer to the character repertoires (as opposed to the legacy encodings) required in Japan market, using terms of ISO/IEC 10646.

2. Proposed collections

The proposed collections are the following five, each of which specification is provided in the last section.

1) BASIC JAPANESE: JIS X 0208 equivalent core subset.

2) JIS2004 IDEOGRAPHICS EXTENSION : JIS X 0213 :2004 Kanji extension.

3) JAPANESE IDEOGRAPHICS SUPPLEMENT : JIS X 0212 :1990 Kanji extension.

4) JAPANESE NON IDEOGRAPHICS EXTENSION : JIS X 0213 :2004 non ideographic extension.

5) COMMON JAPANESE : BASIC JAPANESE with vendor extensions.

3. The intended use of the collections

There are two types of collections in this proposal: core subset and extension/supplement. Both BASIC JAPANESE and COMMON JAPANESE are independent core subsets. Each of the two collections is intended to be used stand-alone or combined with extensions/supplements. Other three collections define additional characters to be combined with a core subset.

Among five collections, BASIC JAPANESE and COMMON JAPANESE define two core subsets. BASIC JAPANESE is basically equivalent to JIS X 0208 and COMMON JAPANESE contains BASIC JAPANESE and widely adopted vendor selected characters including Halfwidth Katakana and Fullwidth Alphanumeric. The two core subsets can be extended in combination with other three collections which are not supposed to be used stand-alone.

The followings are feature descriptions of the three collections for extension:

- JIS2004 IDEOGRAPHICS EXTENISON consists of all level 3 and level 4 Kanji defined in JIS X 0213:2004.
- JAPANESE IDEOGRAPHICS SUPPLEMENT consists of all Kanji defined in JIS X 0212:1990. (Note that 2742 Kanji are common to JIS2004 IDEOGREAPHICS EXTENSION and JAPANESE IDEOGRAPHICS SUPPLEMENT.)
- JAPANESE NON IDEOGRAPHICS EXTENSION is a subset which completes JIS X 0213 non-ideographic repertoire in combination with BASIC JAPANESE or COMMON JAPANESE.

The followings are examples of usage of the proposed collections.
Example-1) BASIC JAPANESE and JIS2004 IDEOGRAPHIC EXTENSION supports JIS X 0208 and level 3 and level 4 Kanji extensions.

Example-2) COMMON JAPANESE + JIS2004 IDEOGRAPHIC EXTENSION + JAPANESE NON IDEOGRAPHIC EXTENSION supports all of JIS X 0213:2004 characters as well as common vendor extensions.

4. Specification of five proposed collections

1) Basic Japanese

Proposed Collection Name: **BASIC JAPANESE**

Number of characters: **6,884**

Collection to be marked as Fixed (Yes / No): **YES**

Code positions in the Collection: **(See BasicJ.txt for code positions except those in CJK UNIFIED IDEOGRAPHS.)**

Collections containing the proposed sub-repertoire:

ID	*UCS-Collection Name / Code Positions*	*Positions to be included or excluded*
1	BASIC LATIN 0020-007E	All code positions are included.
2	LATIN-1 SUPPLEMENT 00A0-00FF	12 code positions listed in BasicJ.txt are included.
8	BASIC GREEK 0370-03CF	48 code positions listed in BasicJ.txt are included.
10	CYRILLIC 0400-04FF	66 code positions listed in BasicJ.txt are included.
32	GENERAL PUNCTUATION 2000-206F	16 code positions listed in BasicJ.txt are included.
36	LETTERLIKE SYMBOLS 2100-214F	Only DEGREE CELSIUS (U+2103) and ANGSTROM SIGN (U+212B) are included.
38	ARROWS 2190-21FF	6 code positions listed in BasicJ.txt are included.
39	MATHEMATICAL OPERATORS 2200-22FF	32 code positions listed in BasicJ.txt are included.
40	MISCELLANEOUS TECHNICAL 2300-23FF	Only ARC (U+2312) is included.

44	BOX DRAWING 2500-257F	32 code positions listed in BasicJ.txt are included.
46	GEOMETRIC SHAPES 25A0-25FF	12 code positions listed in BasicJ.txt are included.
47	MISCELLANEOUS SYMBOLS 2600-26FF	7 code positions listed in BasicJ.txt are included.
49	CJK SYMBOLS AND PUNCTUATION 3000-303F	22 code positions listed in BasicJ.txt are included.
50	HIRAGANA 3040-309F	87 code positions listed in BasicJ.txt are included.
51	KATAKANA 30A0-30FF	90 code positions listed in BasicJ.txt are included.
60	CJK UNIFIED IDEOGRAPHS 4E00-9FFF	6,356 code positions identified by J0 in CJKU_SR.txt (source references for CJK Unified Ideographs) are included.

2) JIS2004 Ideographics Extension

Proposed Collection Name: **JIS2004 IDEOGRAPHICS EXTENSION**

Number of characters: **3,695**

Collection to be marked as Fixed (Yes / No): **YES**

Code positions in the Collection: **(See JIExt.txt)**

Collections containing the proposed sub-repertoire:

ID	***UCS-Collection Name / Code Positions***	***Positions to be included or excluded***
60	CJK UNIFIED IDEOGRAPHS 4E00-9FFF	3,146 code positions listed in JIExt.txt are included.
62	CJK COMPATIBILITY IDEOGRAPHS F900-FAFF	81 code positions listed in JIExt.txt are included.
81	CJK UNIFIED IDEOGRAPHS EXTENSION A 3400-4DBF, FA1F, FA23	165 code positions listed in JIExt.txt are included.
2001	CJK UNIFIED IDEOGRAPHS EXTENSION B 20000-2A6DF	303 code positions listed in JIExt.txt are included.

3) Japanese Ideographics Supplement

Proposed Collection Name: **JAPANESE IDEOGRAPHICS SUPPLEMENT**

Number of characters: **5,801**

Collection to be marked as Fixed (Yes / No): **YES**

Code positions in the Collection:

Collections containing the proposed sub-repertoire:

ID	*UCS-Collection Name / Code Positions*	*Positions to be included or excluded*
60	CJK UNIFIED IDEOGRAPHS 4E00-9FFF	5,801 code positions identified by J1 in CJKU_SR.txt (source references for CJK Unified Ideographs) are included.

4) Japanese Non Ideographics Extension

Proposed Collection Name: **JAPANESE NON IDEOGRAPHICS EXTENSION**

Number of characters: **631**

Collection to be marked as Fixed (Yes / No): **YES**

Code positions in the Collection: **(See JNIExt.txt)**

Collections containing the proposed sub-repertoire:

ID	*UCS-Collection Name / Code Positions*	*Positions to be included or excluded*
2	LATIN-1 SUPPLEMENT 00A0-00FF	83 code positions listed in JNIExt.txt are included.
3	LATIN EXTENDED-A 0100-017F	77 code positions listed in JNIExt.txt are included.
4	LATIN EXTENDED-B 0180-024F	15 code positions listed in JNIExt.txt are included.
5	IPA EXTENSIONS 0250-02AF	55 code positions listed in JNIExt.txt are included.
6	SPACING MODIFIER LETTERS 02B0-02FF	15 code positions listed in JNIExt.txt are included.
7	COMBINING DIACRITICAL MARKS 0300-036F	32 code positions listed in JNIExt.txt are included.
8	BASIC GREEK 0370-03CF	Only GREEK SMALL LETTER FINAL SIGMA (U+03C2) in JNIExt.txt is included.

30	LATIN EXTENDED ADDITIONAL 1E00-1EFF	Only LATIN CAPITAL LETTER M WITH ACUTE (U+1E3E) and LATIN SMALL LETTER M WITH ACUTE (U+1E3F) in JNIExt.txt are included.
31	GREEK EXTENDED 1F00-1FFF	4 code positions listed in JNIExt.txt are included.
32	GENERAL PUNCTUATION 2000-206F	9 code positions listed in JNIExt.txt are included.
34	CURRENCY SYMBOLS 20A0-20CF	Only EURO SIGN (U+20AC) in JNIExt.txt is included.
36	LETTERLIKE SYMBOLS 2100-214F	6 code positions listed in JNIExt.txt are included.
37	NUMBER FORMS 2150-218F	27 code positions listed in JNIExt.txt are included.
38	ARROWS 2190-21FF	10 code positions listed in JNIExt.txt are included.
39	MATHEMATICAL OPERATORS 2200-22FF	23 code positions listed in JNIExt.txt are included.
40	MISCELLANEOUS TECHNICAL 2300-23FF	19 code positions listed in JNIExt.txt are included.
41	CONTROL PICTURES 2400-243F	Only OPEN BOX (U+2423) in JNIExt.txt is included.
43	ENCLOSED ALPHANUMERICS 2460-24FF	66 code positions listed in JNIExt.txt are included.
46	GEOMETRIC SHAPES 25A0-25FF	11 code positions listed in JNIExt.txt are included.
47	MISCELLANEOUS SYMBOLS 2600-26FF	21 code positions listed in JNIExt.txt are included.
48	DINGBATS 2700-27BF	12 code positions listed in JNIExt.txt are included.
49	CJK SYMBOLS AND PUNCTUATION 3000-303F	13 code positions listed in JNIExt.txt are included.
50	HIRAGANA 3040-309F	5 code positions listed in JNIExt.txt are included.
51	KATAKANA 30A0-30FF	6 code positions listed in JNIExt.txt are included.

55	ENCLOSED CJK LETTERS AND MONTHS 3200-32FF	63 code positions listed in JNIExt.txt are included.
56	CJK COMPATIBILITY 3300-33FF	29 code positions listed in JNIExt.txt are included.
66	CJK COMPATIBILITY FORMS FE30-FE4F	Only SESAME DOT (U+FE45) and WHITE SESAME DOT (U+FE46) in JNIExt.txt are included.
69	HALFWIDTH AND FULLWIDTH FORMS FF00-FFEF	Only FULLWIDTH LEFT WHITE PARENTHESIS (U+FF5F) and FULLWIDTH RIGHT WHITE PARENTHESIS (U+FF60) in JNIExt.txt are included.
99	SUPPLEMENTAL ARROWS-B 2900-297F	Only ARROW POINTING RIGHTWARDS THEN CURVING UPWARDS (U+2934) and ARROW POINTING RIGHTWARDS THEN CURVING DOWNWARDS (U+2935) in JNIExt.txt are included.
100	MISCELLANEOUS MATHEMATICAL SYMBOLS-B 2980-29FF	3 code positions listed in JNIExt.txt are included.
102	KATAKANA PHONETIC EXTENSIONS 31F0-31FF	16 code positions listed in JNIExt.txt are included.

5) Common Japanese

Proposed Collection Name: **COMMON JAPANESE**

Number of characters: **7,493**

Collection to be marked as Fixed (Yes / No): **YES**

Code positions in the Collection: **(See CommonJ.txt for code positions except those in BACIC JAPANESE)**

Collections containing the proposed sub-repertoire:

ID	*UCS-Collection Name / Code Positions*	*Positions to be included or excluded*
TBH	BASIC JAPANESE	All code positions included.
32	GENERAL PUNCTUATION 2000-206F	Only HORIZONTAL BAR (U+2015) is included.
36	LETTERLIKE SYMBOLS 2100-214F	Only NUMERO SIGN (U+2116) and

		TELEPHONE SIGN (U+2121) are included.
37	NUMBER FORMS 2150-218F	20 code points defined in JNIExt2.txt are included.
39	MATHEMATICAL OPERATORS 2200-22FF	5 code points defined in JNIExt2.txt are included.
43	ENCLOSED ALPHANUMERICS 2460-24FF	20 code points defined in JNIExt2.txt are included.
49	CJK SYMBOLS AND PUNCTUATION 3000-303F	Only REVERSED DOUBLE PRIME QUOTATION MARK (U+301D) and LOW DOUBLE PRIME QUOTATION MARK (U+301F) are included.
55	ENCLOSED CJK LETTERS AND MONTHS 3200-32FF	8 code points defined in JNIExt2.txt are included.
56	CJK COMPATIBILITY 3300-33FF	28 code points defined in JNIExt2.txt are included.
69	HALFWIDTH AND FULLWIDTH FORMS FF00-FFEF	163 code points defined in JNIExt2.txt are included.
60	CJK UNIFIED IDEOGRAPHS 4E00-9FFF	326 code points defined in JNIExt2.txt are included.
62	CJK COMPATIBILITY IDEOGRAPHS F900-FAFF	32 code points defined in JNIExt2.txt are included.
81	CJK UNIFIED IDEOGRAPHS EXTENSION A 3400-4DBF, FA1F, FA23	Only U+FA1F and U+FA23 are included.

-- end of document --