ISO
INTERNATIONAL ORGANIZATION FOR STANDARDIZATION

ISO/IEC JTC1/SC2/WG2
Universal Multiple-Octet Coded Character Set
(UCS)

X3L2/96-
ISO/IEC JTC1/SC2/WG2 N 1512
Date: January 12, 1997

Title: US contribution for the definition of ISO/IEC 10646 collections
Source: USA (ANSI)
Status: US position
Action: For the consideration of WG2
References: ISO/IEC JTC1/SC2/WG2 N
Distribution: ISO/IEC JTC1/SC2/WG2 members

1. Overview
This document proposes to create a new collection definition reflecting usage as implied by the text of Annex A. The following text is proposed:

4.2 Block: A collection that is contiguous. Blocks do not overlap each other. Characters in a block share common characteristics, such as script.

4.11 Collection: A set of coded characters which is numbered and named and which consists of those coded characters whose code positions lie within one or more identified ranges.

Note: If the identified range includes code positions that are unassigned, the repertoire of the collection will change if additional characters are assigned to one or more of those positions at a future amendment of this standard.

4.12 Specified collection: A collection which contains no code positions reserved for future coding.

2. Rationale
In the original text of ISO/IEC 10646-1: 1993 the normative Annex A 'Collections of graphic characters for subsets' contains an enumeration of numerically identified collections of character encoding ranges. They correspond almost but not exactly to the block names specified in clause 19 of the standard.

The collections as described in Annex A correspond to encoding ranges which contain assigned and unassigned code positions. This allows theses collections to be used through successive standard amendments which add few characters by filling 'holes' in existing ranges. In other words these collections can be used through successive amendments to identify group of characters which belong to a script or a family of symbols. This allows implementations to identify subsets while limiting exposure to amendments.
This applies to the small collections as well as to the BMP as a whole (collection number 300).
In the mean time, an editorial modification has been added to the Standard text providing a definition for the collection as follows: 'A set which is numbered and named and which consists of named characters taken from this standard'
This definition doesn't collect the current understanding of the collection usage, neither does it correspond to the content of Annex A. Therefore the definition of the collection should be amended as proposed in the overview.

Finally the proposed definition of the collection doesn't prevent further definition of a mechanism identifying exact content of the standard at a given time. The concept of a specified collection is proposed to that effect.

The following text hints a how the Annex A of the standard could be reorganized to provide both block information (ordered by range) and collection information (ordered by scripts). This reorganization would remove the need to specify block contents in Clause 19 (Block Names) and concentrate all collection normative content definition in a single location (Annex A).

The first list provides block collections ordered by code point ranges (note that block collections Ids above 100 are new numbers), the second list gives all of the collections arranged by script names:

Blocks

The following table contains the blocks (contiguous, non-overlapping collections) defined in the BMP arranged in sequence of code positions.

Range	Id	Name
0020-007E	1	Basic Latin
00A0-00FF	2	Latin-1 Supplement
0100-017F	3	Latin Extended-A
0180-024F	4	Latin Extended-B
0250-02AF	5	IPA Extensions
02B0-02FF	6	Spacing Modifier Letters
0300-036F	7	Combining Diacritical Marks
0370-03CF	8	Basic Greek
03D0-03FF	9	Greek symbols and Coptic
0400-04FF	10	Cyrillic
0530-058F	11	Armenian
0590-05CF	100	Hebrew Extended-A
05D0-05EA	12	Basic Hebrew
05EB-05FF	101	Hebrew Extended-B
0600-065F	14	Basic Arabic
0660-06FF	15	Arabic Extended
0900-097F	102	Devanagari
0980-09FF	103	Bengali
0A00-0A7F	104	Gurmukhi
0A80-0AFF	105	Gujarati
0B00-0B7F	106	Oriya
0B80-0BFF	107	Tamil
0C00-0C7F	108	Telugu
0C80-0CFF	109	Kannada
0D00-0D7F	110	Malayalam
0E00-0E7F	25	Thai
0E80-0EFF	26	Lao
10A0-10CF	28	Georgian Extended
10D0-10FF	27	Basic Georgian
1100-11FF	29	Hangul Jamo
1E00-1EFF	30	Latin Extended Additional
1F00-1FFF	31	Greek Extended
2000-206F	32	General Punctuation
2070-209F	33	Superscripts and Subscripts
20A0-20CF	34	Currency Symbols
20D0-20FF	35	Combining Diacritical Marks for Symbols
2100-214F	36	Letterlike Symbols
2150-218F	37	Number Forms
2190-21FF	38	Arrows
2200-22FF	39	Mathematical Operators
2300-23FF	40	Miscellaneous Technical
2400-243F	41	Control Pictures
2440-245F	42	Optical Character Recognition

2460-24FF	43	Enclosed Alphanumerics
2500-257F	44	Box Drawing
2580-259F	45	Block Elements
25A0-25FF	46	Geometric Shapes
2600-26FF	47	Miscellaneous Symbols
2700-27BF	48	Dingbats
3000-303F	49	CJK Symbols and Punctuation
3040-309F	50	Hiragana
30A0-30FF	51	Katakana
3100-312F	52	Bopomofo
3130-318F	53	Hangul Compatibility Jamo
3190-319F	54	CJK Miscellaneous
3200-32FF	55	Enclosed CJK Letters and Months
3300-33FF	56	CJK Compatibility
4E00-9FFF	60	CJK Unified Ideographs
E000-F8FF	61	Private Use Area
F900-FAFF	62	CJK Compatibility Ideographs
FB00-FB4F	63	Alphabetic Presentation Forms
FB50-FDFF	64	Arabic Presentation Forms-A
FE20-FE2F	65	Combining Half Marks
FE30-FE4F	66	CJK Compatibility Forms
FE50-FE6F	67	Small Form Variants
FE70-FEFE	68	Arabic Presentation Forms-B
FF00-FFEF	69	Halfwidth and Fullwidth Forms
FEFF-FEFF	70	Specials

Collections

The following list provides collection ordered by script names:

Alphabetical		
7	Combining Diacritical Marks	0300-036F
63	Alphabetic Presentation Forms	FB00-FB4F
Arabic		
14	Basic Arabic	0600-065F
15	Arabic Extended	0660-06FF
64	Arabic Presentation Forms-A	FB50-FDFF
68	Arabic Presentation Forms-B	FE70-FEFE
Armenian		
11	Armenian	0530-058F
Bengali		
17	Bengali	0980-09FF
		200C-200D
Bopomofo		
52	Bopomofo	3100-312F
CJK		
49	CJK Symbols and Punctuation	3000-303F
54	CJK Miscellaneous	3190-319F
55	Enclosed CJK Letters and Months	3200-32FF
56	CJK Compatibility	3300-33FF
60	CJK Unified Ideographs	4E00-9FFF
62	CJK Compatibility Ideographs	F900-FAFF
66	CJK Compatibility Forms	FE30-FE4F
Cyrillic		
10	Cyrillic	0400-04FF
Devanagari		
16	Devanagari	0900-097F
		200C-200D
Georgian		
27	Basic Georgian	10D0-10FF
28	Georgian Extended	10A0-10CF
Greek		

8	Basic Greek	0370-03CF
9	Greek symbols and Coptic	03D0-03FF
31	Greek Extended	1F00-1FFF
Gurmukhi		
18	Gurmukhi	0A00-0A7F
		200C-200D
Gujarati		
19	Gujarati	0A80-0AFF
		200C-200D
Hangul		
29	Hangul Jamo	1100-11FF
53	Hangul Compatibility Jamo	3130-318F
??	Hangul Syllables	AC00-D7A3
Hebrew		
12	Basic Hebrew	05D0-05EA
13	Hebrew Extended	0590-05CF
		05EB-05FF
Kana		
50	Hiragana	3040-309F
51	Katakana	30A0-30FF
Kannada		
23	Kannada	0C80-0CFF
		200C-200D
Lao		
26	Lao	0E80-0EFF
Latin		
1	Basic Latin	0020-007E
2	Latin-1 Supplement	00A0-00FF
3	Latin Extended-A	0100-017F
4	Latin Extended-B	0180-024F
5	IPA Extensions	0250-02AF
6	Spacing Modifier Letters	02B0-02FF
30	Latin Extended Additional	1E00-1EFF
Malayalam		
24	Malayalam	0D00-0D7F
		200C-200D
Miscellaneous		
61	Private Use Area	E000-F8FF
65	Combining Half Marks	FE20-FE2F
67	Small Form Variants	FE50-FE6F
69	Halfwidth and Fullwidth Forms	FF00-FFEF
70	Specials	FEFF-FEFF
??	High Surrogates	D800-DB7F
??	High Private Use Surrogates	DB80-DBFF
??	Low Surrogates	DC00-DFFF
Oriya		
20	Oriya	0B00-0B7F
		200C-200D
Symbols		
32	General Punctuation	2000-206F
33	Superscripts and Subscripts	2070-209F
34	Currency Symbols	20A0-20CF
35	Combining Marks for Symbols	20D0-20FF
36	Letterlike Symbols	2100-214F
37	Number Forms	2150-218F
38	Arrows	2190-21FF
39	Mathematical Operators	2200-22FF
40	Miscellaneous Technical	2300-23FF
41	Control Pictures	2400-243F
42	Optical Character Recognition	2440-245F
43	Enclosed Alphanumerics	2460-24FF

44	Box Drawing	2500-257F
45	Block Elements	2580-259F
46	Geometric Shapes	25A0-25FF
47	Miscellaneous Symbols	2600-26FF
48	Dingbats	2700-27BF
Tamil		
21	Tamil	0B80-0BFF
		200C-200D
Telugu		
22	Telugu	0C00-0C7F
		200C-200D
Thai		
25	Thai	0E00-0E7F
Tibetan		
??	Tibetan	0F00-0FBF

[end of US contribution]