X11/96-42

MUMPS Development Committee

Extension to the MDC Standard
Type A Release of the MUMPS Development Committee

charset: ISO-8859-1-USA

March 23, 1996

Produced by the MDC Subcommittee #12
Environment

Ed de Moel, Chairman
MUMPS Development Committee

Larry Ruh, Chairman
Subcommittee #12

The reader is hereby notified that the following MDC specification has been approved by the MUMPS Development Committee but that it may be a partial specification which relies on information appearing in many parts of the MDC Standard. This specification is dynamic in nature, and the changes reflected by this approved change may not correspond with the latest specification available.

Because of the evolutionary nature of MDC specifications, the reader is further reminded that changes are likely to occur in the specification released herein prior to a complete republication of the MDC Standard.

1. Identification

1.1 Title:

charset: ISO-8859-1-USA

1.2 MDC Proposer and Sponsor:

Proposer:
Ben Bishop
64 Maolis Road
Nahant, MA 01908
(617) 593-3038
aci@shore.net

Sponsor:
SC12/TG2 Internationalisation
Kate Schell, Chair
Jaquard Systems Research
(617) 646-9033
cschell@world.std.com

1.3 Motion:

None. Final publication version superseding X11/SC12/96-3.

1.4 History:

Date	Document	Action	
01 Aug 96	X11/96-42	Final publication version.	
01 Feb 96	X11/SC12/96-3	Proposed as MDC/A	Passed 23:0:4
31 Aug 95	X11/SC12/TG2/95-5	Proposed as SC12/A	Passed 11:1:4
01 Jul 95	X11/SC12/TG2/95-2	re-split into collation statement & definition	
04 Jun 95	X11/SC12/TG2/95-2	Split, proposed as SC12/B	Passed 5:0:2
19 Apr 95	X11/SC12/TG2/95-2	Proposed as SC12/B Revised	
01 Dec 94	X11/SC12/TG2/94-7	Initial proposal SC12/B	Passed 4:0:2

1.5 Dependencies:

No proposals have been identified which depend on this proposal.
This document uses the collation ordering definition from "String and M Collation" (X11/96-41).

2. Justification

2.1 Needs

The reality of internationalisation (i18n) goes far beyond the need to incorporate non-ASCII characters within programs and data. The i18n proposals provided for the syntactic elements and procedures for incorporating international character sets into the MUMPS environment. One of the most likely candidates for inclusion as a character set profile is ISO-8859-1, also known as ISO-Latin-1.

Due to cultural bias in collation rules (especially those in countries using the ISO-8859-1 character set), it is important to narrow the application of the character set profile. Although two languages may use the same alphabet/character set, the collation rule, or even patcode values might be different. This proposal addresses the ISO-8859-1 ordering and pattern matching within the context of the United States, and representing foreign words or characters within that context.

An additional, sibling, character set profile will also be defined, called ISO-8859-1-USA/M. This character set profile is identical to that of ISO-8859-1-USA except that canonic numbers collate first. The intention is to mimic the relationship between the ASCII and M Character set profiles.

2.2 Existing Practice

Several vendors support inclusion of non-ASCII data, but none (to the author's knowledge) have implemented a version of the i18n proposals for a non-ASCII based (i.e. 8-bit) character set. Note also that OMI (X11.2) uses the equivalent of the ISO-8859-1, but does not define a collation order other than the strict interpretation of the character codes (code order).

3. Description

3.1 General description

Build a character set definition table: use the character set ISO-8859-1 which contains 256 characters; each row of the definition table represents a character defined by its ID (a decimal number in the range 0-255), the common symbol it represents in ISO-8859-1 (if available), the patcodes it matches, as well as the list of collation ordering information needed to properly collate the information. This collation information is designed to accomplish a multi-pass ordering as described in SC22/WG20 N221 Draft 3, October 1993: European Multilingual Ordering). The passes have a hierarchical order; if some pass cannot determine precedence of one of the two strings, then the next pass is taken. During the first pass, the shorter of two equivalent strings is considered 'first'. (The following is a paraphrase of the relevant passes described in SC22/WG20 N221)

i. Digits are ordered before letters. Letters with diacritical marks are considered equivalent to the corresponding letters without diacritical marks. Lowercase and corresponding upper case letters are considered equivalent. A number of strings of characters (including "ligatures") are equivalenced to strings of letters (see below).

ii. All letters without diacritical marks are considered equivalent. All other letters with a specific diacritical mark or with a specific combination of diacritical marks are considered equivalent. The diacritical marks are ordered according to ISO 6937: acute, grave, circumflex, diaeresis, tilde, caron, breve, double acute, ring above, dot above, macron, cedilla, ogonek, dotless, stroke.

iii. All SMALL letters form an equivalence class <SMALL> and collate first; all CAPITAL letters form an equivalence class <CAPITAL> and collate next, and all other letters form an equivalence class of <OTHER>. Digits are ordered into the equivalence classes <NORMAL-DIGIT>, <SUPERSCRIPT>, <SUBSCRIPT> and <OTHER>. All other characters are ignored.

iv. The collation sequence of the characters is their order of appearance in ISO/IEC 10646-1.

As applied to ISO-8859-1 within the context of United States collation, the comparison is done in passes until an ordering is determined; this has the following behavior:

- 1st order collation values are compared. The supplied values cause digits to collate first, in addition, upper & lower case forms of letters collate as if they were the same and as if they did not have diacritic marks (i.e. Ä, A, â, and Á all 1st-order collate as if they were all the same character)

- 2nd order collation values are compared. Lowercase versions of letters (still ignoring diacritic marks) are ordered first, uppercase second.

- 3rd order collation values are compared. Ordering is based on the diacritic marks of the letters, in order according to ISO 6937 (see above). Note that ligatures (single characters which are compared as if they were specific strings) are ordered based on an implied "diacritical" mark which causes them to order before their equivalent string. ISO-8859-1 contains three ligatures:

  - Æ (id# 198) is collated "as if" it were "AE"
  - æ (id# 230) is collated "as if" it were "ae"
  - ß (id# 223) is collated "as if" it were "ss"

- Finally (if all other passes are equal), the actual id values of the characters are compared. No two non-equal strings will have the same id values for the constituent characters.

3.2 Annotated Examples of Use

Unavailable at the present time; should this be examples of what the collation order *should* be for a selected set of ISO-8859-1 characters?

3.3 Formalization (References are to the X11.1-1994 Canvass Document)

- **Add to Annex A (Normative) the following two charset definitions:**

charset ISO-8859-USA

The charset ISO-8859-1-USA is defined using the table A.2. The values in the columns headed Character ID and Character Symbol are taken from ISO-8859-1 (ISO Latin 1). The column headed patcode defines which characters match the patcodes A, C, E, I, L, N, P, and U. The characters in the table with a patcode of A are defined as idents. The collation rule used is *STRING* collation, using the collation order values provided in the table: note that all collation is left-to-right precedence. Note also that the patcode I matches any non-ASCII characters (id# greater than 127), not just those listed in this charset.

charset ISO-8859-1-USA/M

The charset ISO-8859-1-USA is defined using the table A.2. The values in the columns headed Character ID and Character Symbol are taken from ISO-8859-1 (ISO Latin 1). The column headed patcode defines which characters match the patcodes A, C, E, I, L, N, P, and U. The characters in the table with a patcode of A are defined as idents. The collation rule used is *M* collation, using the collation order values provided in the table: note that all collation is left-to-right precedence. Note also that the patcode I matches any non-ASCII characters (id# greater than 127), not just those listed in this charset.

4. Implementation Effects

4.1 Effect on Existing User Practices and Investments

None expected.

4.2 Effect on Existing Vendor Practices and Investments

Costs of implementation. Note that this proposed charset uses the existing, standard, structures for using a non-ASCII character set, specifically ^$CHARACTER)

4.3 Techniques and Costs for Compliance Verification

Test the aspects of the charset: make sure all the characters match the appropriate patcodes. Confirm the single-character collation sequence is as expected; then generate two and three character sequences to confirm the ligature and diacritical character collation sequences as expected. Also test the *M* collation version using canonic and non-canonic numbers and how they sort compared to non-numeric strings.

4.4 Legal Considerations

None identified.

5. Closely Related Standards Activities

5.1 Other X11 Proposals Under Consideration

None.

5.2 Other Related Standards Efforts

None.

5.3 Recommendations for Coordinating Liaison

None.

6. Associated Documents

X11.1-1995 Standard, Annex A	ASCII, M and JIS90 charsets
X11/95-15	charset names
X11/96-41	String & M collation
ISO-6937	Diacritical mark ordering
ISO 8859-1:1987	8 Bit single-byte Coded Graphic Character Sets - Part 1: Latin Alphabet No. 1
ISO/SC22/WG20 N221	European Multilingual Ordering

7. Issues, Pros and Cons, and Discussion

- January 1995: Albuquerque, NM: Proposed as SC12/B, passed 4:0:2
  Pros: allows structured definition of charset names. Allows for standardized usage of ISO 8859-1 within M applications. Cons: Mixes two amendments in one proposal. Profile may not be completely accurate.
  In answer to con 1: rather than make a pair of co-dependant proposals, the author feels that a single proposal is more appropriate since this is the only proposal trying to use 'descriptor'.
  In answer to con 2: this proposal has a modified form of the collation algorithm which should allow arbitrary collation definitions which are flexible enough for other character sets in the future.

- June 1995: Chicago, IL: broken into separate proposals (charset names, ISO-8859-1-USA)
  ISO-8859-1-USA was revised; "charset names" was approved as SC12/A and collation ordering was modified and separated out. This new version only deals with the ISO-8859-1-USA table and definition. Complaints about certain orderings were raised which the author hopes have been addressed. One specific complaint was the symbol mu (μ) was defined as a punctuation rather than a lowercase letter (greek); the reason for this is that ISO-8859-1 is not intended to represent the greek alphabet -- ISO-8859-2 is designed for that. So, in the context of ISO-8859-1, mu is a punctuation mark.

- July 1995: Nahant, MA: re-split into two proposals: (String and M Collation and ISO-8859-1)
  Upon reading the 'splitting' of the original proposal in the minutes, I felt it was also advisable to separate out the charset definition part of the ISO-8859-1 from the reformulazation of the collation rules for String and M Collation.

- October 1995: New Orleans, LA: Proposed as SC12/A, passed 11:1:4
  Pro: 1. allows for usage of ISO 8859-1 USA
  2. provides a model of a character set profile
  3. puts western europe back on an equal footing with Japan
  4. NOT con 1: The character set for portable M code is the M charset.
  Con: 1. Breaks existing code.

- March 1995: Boston, MA: Proposed as MDC/A, passed 23:0:4
  Pro: 1. Allows usage of ISO-8859-1-USA
  2. Provides model for a [non-trivial] character set profile
  3. Character set for portable M code is M
  Con: 1. Breaks existing code
  Response to con: The idea that adding a charset would break existing code carries no weight. The argument was that programmers would use ISO-8859-1/USA as their default charset and therefore the routines would not be portable, ergo breaking those routines. The answering argument (Pro 3) was that the 'official' character set for portable M code is the M charset; there are rules to be followed if a routine is expected to be portable, and this is an existing (but mostly moot to present) rule.

8. Glossary (taken from document SC22/WG20/N312 // ISO/WD 12 199.3)

character (1) A member of a set of elements used for the organisation, control, or representation of data.

(2) A character is a simple or composite graphic symbol belonging to a conventional set of symbols. There are alphabetic characters, numerical characters (arabic and roman), diacritic characters (for example ^ ° ' ‘), punctuation characters (for example . , ; : ! ?) and specific other characters (for example § $ % & { #). - Synonyms which should be avoided: {graphic, phonetic} symbol, sign, mark, note, cipher.

combining character: A member of an identified subset of the coded character set of ISO/IEC 10646 intended for combination with the preceding non-combining graphic character, or with

a sequence of combining characters preceded by a non-combining character (see also *composite sequence*). NOTE - This part of ISO/IEC 10646 specifies several subset collections which include combining characters.

composite sequence: A sequence of graphic characters consisting of a non-combining character followed by one or more combining characters (see also *combining characters*). NOTES - 1:A graphic symbol for a composite sequence generally consists of the combination of the graphic symbols of each character in the sequence - 2: A composite sequence is not a character and therefore is not a member of the repertoire of ISO/IEC 10646.

default state: The state that is assumed when no state has been explicitly specified.

diacritic: Character which is not a [letter] of the latin alphabet and which is placed over, under, or through a letter or a combination of letters indicating a semantic or phonetic value different from that given the unmarked or otherwise marked letter. A letter with a diacritic is a composite character. NOTE - The point of german "umlaut"-character should be regarded as diacritic. [also called a *diacritical mark*]

diacritic mark: An attribute used of a character applied to denote a variation of a letter.

digit: A graphic character used to represent the numeric value, or part thereof, of a number. Examples: decimal digits, hexadecimal digits.

graphic character: A character, other than a control function, that has a visual representation normally handwritten, printed, or displayed.

letter: (1) A letter (or alphabetic character) is a character which is an individual unspecific basic unit of the latin alphabet, irrespective of the shape and any graphical realization on a medium. A letter can be specified as a *small letter* or *capital letter*.

(2) A graphic character used for writing natural language, normally representing a sound of the language.

ligature: (1) A composite character joining two or more letters. There are ligatures which are conventionalized units of a national variant of the latin alphabet, and ligatures which are caused by the font used in a document. Maybe the first ones should be named *ligature characters*, the last ones *ligature font elements*. [Language dependent. Only ligature characters are taken into consideration]

(2) Two or more letters written together. The resulting symbol is in some cases considered equivalent with the originating letters, in some cases it is considered a separate entity.

ordering: Bringing strings of characters into a well-defined sequence using a string comparison specification.

9. Appendix

Appendix A: ISO-8859-1 Character set, with USA collation list and pattern match character designations.

Table A.2 - ISO-8859-1-USA Character Set Table

Character ID	Character Symbol	patcode	Collation Table		
			1st Order	2nd Order	3rd Order
0	*NUL*	C,E	0		
1	*SOH*	C,E	1		
2	*STX*	C,E	2		
3	*ETX*	C,E	3		
4	*EOT*	C,E	4		
5	*ENQ*	C,E	5		
6	*ACK*	C,E	6		
7	*BELL*	C,E	7		
8	*BS*	C,E	8		
9	*HT*	C,E	9		
10	*LF*	C,E	10		
11	*VT*	C,E	11		
12	*FF*	C,E	12		
13	*CR*	C,E	13		
14	*SO*	C,E	14		
15	*SI*	C,E	15		
16	*DLE*	C,E	16		
17	*DC1*	C,E	17		
18	*DC2*	C,E	18		
19	*DC3*	C,E	19		
20	*DC4*	C,E	20		
21	*NAK*	C,E	21		
22	*SYN*	C,E	22		
23	*ETB*	C,E	23		
24	*CAN*	C,E	24		

Character ID	Character Symbol	patcode	Collation Table		
			1st Order	2nd Order	3rd Order
25	*EM*	C,E	25		
26	*SUB*	C,E	26		
27	*ESC*	C,E	27		
28	*FS*	C,E	28		
29	*GS*	C,E	29		
30	*RS*	C,E	30		
31	*US*	C,E	31		
32	*SP*	P,E	32		
33	!	P,E	33		
34	"	P,E	34		
35	#	P,E	35		
36	$	P,E	36		
37	%	P,E	37		
38	&	P,E	38		
39	'	P,E	39		
40	(	P,E	40		
41	)	P,E	41		
42	*	P,E	42		
43	+	P,E	43		
44	,	P,E	44		
45	-	P,E	45		
46	.	P,E	46		
47	/	P,E	47		
48	0	N,E	48		
49	1	N,E	49		
50	2	N,E	50		
51	3	N,E	51		

Character ID	Character Symbol	patcode	Collation Table		
			1st Order	2nd Order	3rd Order
52	4	N,E	52		
53	5	N,E	53		
54	6	N,E	54		
55	7	N,E	55		
56	8	N,E	56		
57	9	N,E	57		
58	:	P,E	58		
59	;	P,E	59		
60	<	P,E	60		
61	=	P,E	61		
62	>	P,E	62		
63	?	P,E	63		
64	@	P,E	64		
65	A	A,U,E	65	1	1
66	B	A,U,E	66	1	1
67	C	A,U,E	67	1	1
68	D	A,U,E	68	1	1
69	E	A,U,E	70	1	1
70	F	A,U,E	71	1	1
71	G	A,U,E	72	1	1
72	H	A,U,E	73	1	1
73	I	A,U,E	74	1	1
74	J	A,U,E	75	1	1
75	K	A,U,E	76	1	1
76	L	A,U,E	77	1	1
77	M	A,U,E	78	1	1
78	N	A,U,E	79	1	1

Character ID	Character Symbol	patcode	Collation Table		
			1st Order	2nd Order	3rd Order
79	O	A,U,E	80	1	1
80	P	A,U,E	81	1	1
81	Q	A,U,E	82	1	1
82	R	A,U,E	83	1	1
83	S	A,U,E	84	1	1
84	T	A,U,E	85	1	1
85	U	A,U,E	86	1	1
86	V	A,U,E	87	1	1
87	W	A,U,E	88	1	1
88	X	A,U,E	89	1	1
89	Y	A,U,E	90	1	1
90	Z	A,U,E	91	1	1
91	[	P,E	93		
92	\	P,E	94		
93	]	P,E	95		
94	^	P,E	96		
95	_	P,E	97		
96	‘	P,E	98		
97	a	A,L,E	65	0	1
98	b	A,L,E	66	0	1
99	c	A,L,E	67	0	1
100	d	A,L,E	68	0	1
101	e	A,L,E	70	0	1
102	f	A,L,E	71	0	1
103	g	A,L,E	72	0	1
104	h	A,L,E	73	0	1
105	i	A,L,E	74	0	1

Character ID	Character Symbol	patcode	Collation Table		
			1st Order	2nd Order	3rd Order
106	j	A,L,E	75	0	1
107	k	A,L,E	76	0	1
108	l	A,L,E	77	0	1
109	m	A,L,E	78	0	1
110	n	A,L,E	79	0	1
111	o	A,L,E	80	0	1
112	p	A,L,E	81	0	1
113	q	A,L,E	82	0	1
114	r	A,L,E	83	0	1
115	s	A,L,E	84	0	1
116	t	A,L,E	85	0	1
117	u	A,L,E	86	0	1
118	v	A,L,E	87	0	1
119	w	A,L,E	88	0	1
120	x	A,L,E	89	0	1
121	y	A,L,E	90	0	1
122	z	A,L,E	91	0	1
123	{	P,E	99		
124	|	P,E	100		
125	}	P,E	101		
126	~	P,E	102		
127	*DEL*	C,E	103		
128		C,E,I	104		
129		C,E,I	105		
130		C,E,I	106		
131		C,E,I	107		
132	*IND*	C,E,I	108		

Character ID	Character Symbol	patcode	Collation Table		
			1st Order	2nd Order	3rd Order
133	*NEL*	C,E,I	109		
134	*SSA*	C,E,I	110		
135	*ESA*	C,E,I	111		
136	*HTS*	C,E,I	112		
137	*HTJ*	C,E,I	113		
138	*VTS*	C,E,I	114		
139	*PLD*	C,E,I	115		
140	*PLU*	C,E,I	116		
141	*RI*	C,E,I	117		
142	*SS2*	C,E,I	118		
143	*SS3*	C,E,I	119		
144	*DCS*	C,E,I	120		
145	*PU1*	C,E,I	121		
146	*PU2*	C,E,I	122		
147	*STS*	C,E,I	123		
148	*CCH*	C,E,I	124		
149	*MW*	C,E,I	125		
150	*SPA*	C,E,I	126		
151	*EPA*	C,E,I	127		
152		C,E,I	128		
153		C,E,I	129		
154		C,E,I	130		
155	*CSI*	C,E,I	131		
156	*ST*	C,E,I	132		
157	*OSC*	C,E,I	133		
158	*PM*	C,E,I	134		
159	*APC*	C,E,I	135		

Character ID	Character Symbol	patcode	Collation Table		
			1st Order	2nd Order	3rd Order
160	*NBSP*	C,E,I	136		
161	¡	P,E,I	137		
162	¢	P,E,I	138		
163	£	P,E,I	139		
164	¤	P,E,I	140		
165	¥	P,E,I	141		
166	¦	P,E,I	142		
167	§	P,E,I	143		
168	¨	P,E,I	144		
169	©	P,E,I	145		
170	ª	P,E,I	146		
171	«	P,E,I	147		
172	¬	P,E,I	148		
173	—	P,E,I	149		
174	®	P,E,I	150		
175	¯	P,E,I	151		
176	°	P,E,I	152		
177	±	P,E,I	153		
178	²	P,E,I	154		
179	³	P,E,I	155		
180	´	P,E,I	156		
181	µ	P,E,I	157		
182	¶	P,E,I	158		
183	·	P,E,I	159		
184	¸	P,E,I	160		
185	¹	P,E,I	161		
186	º	P,E,I	162		

Character ID	Character Symbol	patcode	Collation Table		
			1st Order	2nd Order	3rd Order
187	»	P,E,I	163		
188	¼	P,E,I	164		
189	½	P,E,I	165		
190	¾	P,E,I	166		
191	¿	P,E,I	167		
192	À	A,U,E,I	65	1	3
193	Á	A,U,E,I	65	1	2
194	Â	A,U,E,I	65	1	4
195	Ã	A,U,E,I	65	1	6
196	Ä	A,U,E,I	65	1	5
197	Å	A,U,E,I	65	1	10
198	Æ	A,U,E,I	65 70	1 1	1 0
199	Ç	A,U,E,I	67	1	13
200	È	A,U,E,I	70	1	3
201	É	A,U,E,I	70	1	2
202	Ê	A,U,E,I	70	1	4
203	Ë	A,U,E,I	70	1	5
204	Ì	A,U,E,I	74	1	3
205	Í	A,U,E,I	74	1	2
206	Î	A,U,E,I	74	1	4
207	Ï	A,U,E,I	74	1	5
208	Ð	A,U,E,I	69	1	1
209	Ñ	A,U,E,I	79	1	6
210	Ò	A,U,E,I	80	1	3
211	Ó	A,U,E,I	80	1	2
212	Ô	A,U,E,I	80	1	4

Character ID	Character Symbol	patcode	Collation Table		
			1st Order	2nd Order	3rd Order
213	Õ	A,U,E,I	80	1	6
214	Ö	A,U,E,I	80	1	5
215	×	P,E,I	168		
216	Ø	A,U,E,I	80	1	16
217	Ù	A,U,E,I	86	1	3
218	Ú	A,U,E,I	86	1	2
219	Û	A,U,E,I	86	1	4
220	Ü	A,U,E,I	86	1	5
221	Ý	A,U,E,I	90	1	2
222	þ	A,U,E,I	92	1	1
223	ß	A,L,E,I	84 84	0 0	1 0
224	à	A,L,E,I	65	0	3
225	á	A,L,E,I	65	0	2
226	â	A,L,E,I	65	0	4
227	ã	A,L,E,I	65	0	6
228	ä	A,L,E,I	65	0	5
229	å	A,L,E,I	65	0	10
230	æ	A,L,E,I	65 70	0 0	1 0
231	ç	A,L,E,I	67	0	13
232	è	A,L,E,I	70	0	3
233	é	A,L,E,I	70	0	2
234	ê	A,L,E,I	70	0	4
235	ë	A,L,E,I	70	0	5
236	ì	A,L,E,I	74	0	3
237	í	A,L,E,I	74	0	2

Character ID	Character Symbol	patcode	Collation Table		
			1st Order	2nd Order	3rd Order
238	î	A,L,E,I	74	0	4
239	ï	A,L,E,I	74	0	5
240	ð	A,L,E,I	69	0	1
241	ñ	A,L,E,I	79	0	6
242	ò	A,L,E,I	80	0	3
243	ó	A,L,E,I	80	0	2
244	ô	A,L,E,I	80	0	4
245	õ	A,L,E,I	80	0	6
246	ö	A,L,E,I	80	0	5
247	÷	P,E,I	169		
248	ø	A,L,E,I	80	0	16
249	ù	A,L,E,I	86	0	3
250	ú	A,L,E,I	86	0	2
251	û	A,L,E,I	86	0	4
252	ü	A,L,E,I	86	0	5
253	ý	A,L,E,I	90	0	2
254	þ	A,L,E,I	92	0	1
255	ÿ	A,L,E,I	90	0	5

Note: unique collation requires that no two rows of this table have identical collation order columns.