X11/96-41

MUMPS Development Committee

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

String and *M* Collation

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:

String and *M* Collation

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
C Schell Systems
(617) 646-9033
cschell@world.std.com

1.3 Motion:

None. Final Publication version; superseding X11/SC12/96-2.

1.4 History:

Date	Document	Action	
01 Aug 96	X11/96-41	Final Publication Version	
01 Feb 96	X11/SC12/96-2	Proposed as MDC/A:	Passed: 23-0-4
31 Aug 95	X11/SC12/TG2/95-4	Proposed as SC12/A:	Passed: 12-0-5
01 Jul 95	X11/SC12/TG2/95-2	re-split of ISO-8859-1 into collation statement & definition	
04 Jun 95	X11/SC12/TG2/95-2	ISO-8859-1 proposal split into 2 parts (names and body)	
19 Apr 95	X11/SC12/TG2/95-2	Proposed as SC12/B:	Passed: 5-0-2
01 Dec 95	X11/SC12/TG2/94-7	Initial ISO-8859-1-USA proposal as SC12/B	Passed: 4-0-2

1.5 Dependencies:

The document X11/96-42 (charset: ISO-8859-1-USA) uses the collation definition created by this proposal.

No proposals have been identified upon which this proposal depends.

2. Justification

2.1 Needs

Upon examination of Annex A of the X11.1 Canvass Document, each charset appears expected to detail the algorithm by which collation is determined; specifically, both the ASCII and M charsets define their collation ordering function (CO). A definition is needed which is generic and applicable to more than just the ASCII character set; the differences should all be contained in the table associated with the character set, much in the same way that the patcodes and character IDs are presented. Note that this does not prevent other charset definitions from providing their own, custom, definitions where this schema does not offer the flexibility needed to accurately represent the charset

2.2 Existing Practice

Each charset definition needs to have the Collation Ordering function described verbosely.

3. Description

3.1 General description

Define 'String' and 'M' collation; this involves describing the format of the Character Set table so that it includes Collation Ordering columns with both an expected precedence and collation ordering (i.e. left to right or right to left). Character Set tables presented in this format may then use these 'standardized' collation rules rather than re-writing the existing rules to apply to individual character sets.

3.2 Annotated Examples of Use

See Annex A for the presentation of the ASCII character set using this table format; this Annex is intended to replace paragraph I and II of Annex A.

3.3 Formalization

- **Annex A: insert after the paragraph beginning "The definition of a Character Set Profile ...":**

Note that the patcodes A, C, E, L, N, P, and U are applicable for all character set profiles; in addition patcode E matches any character, not just those listed in any specific charset.

Two collation schemes are provided which only require a properly defined table of characters for the Character Set associated with the specific Character Set Profile.

STRING COLLATION

Determining the Collation Ordering for a Character Set Profile requires the collation value(s) for each character within the character set be accessible as a group of values presented as an *n*-tuple. Each column of the definition table provides one value of the tuple in the specified order. When no value is present in any column, the corresponding character id value is used in its place. Note that certain characters may be represented with more than one value entry line in the table; in these cases the entries are taken one at a time and treated as if they represented separate characters in the original string (e.g. the character Æ in ISO-Latin-1 (id# 198) would be treated as a form of the string "AE").

Let *s* be any non-empty string. Define the numeric function $CV_n(s)$ to return the nth-order collation value for string *s*: unless otherwise specified this value is determined by evaluating the value in the *n*th-column of each collation tuple for each character in the string examined in left-to-right order and combining those tuples together. Note: selected collation-tuple columns may optionally be designated for right-to-left evaluation.

The Collation Ordering function CO determines relative ordering for a character set. The exact value of this function is not specified here, however, the values formed by any implementation must satisfy the following rules when comparing two non-equal strings:

Let t also be any non-empty string, not equal to s. The *STRING* Collation Ordering function CO is defined as:

a) $CO("", s) = s$

b) $CO(s, t) = t$ if, and only if, $CV_j(t) > CV_j(s)$ and for all i, i=1 ... j-1, $CV_i(t) = CV_i(s)$; otherwise $CO(s,t) = s$.

M COLLATION

The *M* Collation Ordering function CO for uses the definition of $CV_n(s)$ specified in *STRING* Collation and is otherwise different only with respect to numbers:

Let s be any non-empty string, let m and n be strings satisfying the definition of numeric data values (see I.7.1.4.3), and u and v be non-empty strings which do not satisfy that definition.

a) $CO("", s) = s$

b) $CO(m, n) = n$ if $n>m$; otherwise, $CO(m, n) = m$

c) $CO(m, u) = u$

d) $CO(u, v) = v$ if, and only if, $CV_j(v) > CV_j(u)$ and for all i, i=1 ... j-1, $CV_i(v) = CV_i(u)$; otherwise, $CO(u, v) = u$.

- **Replace numbered sections 1 & 2 (charset M and charset ASCII) in Annex A with the following:**

1 charset M .

The charset M is defined using the table A.1. The values in the columns headed Character ID and Character Symbol are taken from ASCII (X3.4-1990). The column headed patcode defines which characters match the patcodes A, C, E, 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.

2. charset ASCII

The charset ASCII is defined using the table A.1. The values in the columns headed Character ID and Character Symbol are taken from ASCII (X3.4-1990). The column headed patcode defines which characters match the patcodes A, C, E, 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.

- **Insert after the (new) numbered section 2 in Annex A, the contents of this proposal's Appendix A.**

- **Insert at an appropriate place in Part II (portability section) of the standard (with a new header as needed):**

  Collation values are not portable between implementations unless the value is explicitly stated in the definition of the Character Set Profile.

4. Implementation Effects

4.1 Effect on Existing User Practices and Investments

None expected.

4.2 Effect on Existing Vendor Practices and Investments

None expected.

4.3 Techniques and Costs for Compliance Verification

None.

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 1994 canvass document

7. Issues, Pros and Cons, and Discussion

Defining the collation rules in a generic manner will permit future character set profiles to be presented without having to "reinvent the wheel" each time. Note that this proposal does not change either the M or ASCII character set profiles in any substantial way, it just changes the ways in which they are defined.

- October 1995: New Orleans, LA proposed as SC12/A, passed 12:0:5
  At the June 1995 meeting (Chicago), the ISO-8859-1-USA proposal was broken into two separate proposals, charset names (now MDC/A) and the ISO-8859-1-USA proposal body (the names segment of the proposal was distinct and available to go forward). In July I split the remaining ISO-8859-1-USA proposal in two again, consisting of the definition of charsets and their collation rules, and the ISO-8859-1-USA character set profile since once again they were distinct and would permit refinement of ISO-8859-1 without holding up the charset definition (this document).
  There was 1 pro, no cons; pro: Simplifies charset definition.

- March 1996, Boston, MA proposed as MDC/A, passed 23-0-4
  Pro: Simplifies charset definition.

8. Glossary

None.

9. Appendix

A. ASCII & M character set definition for Annex A table A.1

Table A.1 - ASCII 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 (space)	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	' (apostrophe)	P,E	39		
40	(	P,E	40		
41	)	P,E	41		
42	*	P,E	42		
43	+	P,E	43		
44	, (comma)	P,E	44		
45	- (hyphen)	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		
66	B	A,U,E	66		
67	C	A,U,E	67		
68	D	A,U,E	68		
69	E	A,U,E	69		
70	F	A,U,E	70		
71	G	A,U,E	71		
72	H	A,U,E	72		
73	I	A,U,E	73		
74	J	A,U,E	74		
75	K	A,U,E	75		
76	L	A,U,E	76		
77	M	A,U,E	77		
78	N	A,U,E	78		

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

Character ID	Character Symbol	patcode	Collation Table		
			1st Order	2nd Order	3rd Order
106	j	A,L,E	106		
107	k	A,L,E	107		
108	l	A,L,E	108		
109	m	A,L,E	109		
110	n	A,L,E	110		
111	o	A,L,E	111		
112	p	A,L,E	112		
113	q	A,L,E	113		
114	r	A,L,E	114		
115	s	A,L,E	115		
116	t	A,L,E	116		
117	u	A,L,E	117		
118	v	A,L,E	118		
119	w	A,L,E	119		
120	x	A,L,E	120		
121	y	A,L,E	121		
122	z	A,L,E	122		
123	{	P,E	123		
124	|	P,E	124		
125	}	P,E	125		
126	~	P,E	126		
127	*DEL*	C,E	127		

Note: 2nd and 3rd order collation values happen to be blank (i.e. not needed) for this Character Set Profile definition; the 1st order collation value happens to be unique across all the characters in this profile.