Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation Internationale de Normalisation
Международная организация по стандартизации

Doc Type:	**Working Group Document**
Title:	**Proposal to encode segmented hexadecimal digits from Chinese legacy operation system**
Source:	**Eiso Chan (陈永聪)**
Status:	**Individual Contribution to UTC**
Action:	**For consideration by UTC**
Date:	**2022-05-14**

0. Introduction

Doug Ewell, Rebecca Bettencourt, Ricardo Bánffy, Michael Everson, Eduardo Marín Silva, Elias Mårtenson, Mark Shoulson, Shawn Steele and Rebecca Turner prosed to add the characters from legacy computers and teletext in L2/19-025 and provided the huge mappings. In the source table, they showed 10 segmented digits were cited from *Atari ST Character Set*. These 10 digits just supported the decimal digit representations.

CCDOS (*Chinese Characters Disk Operation System*, please see the Chinese introduction here) was developed by the Sixth Research Institute of the Ministry of Electronic Industry of PRC in 1980s, which once was the most common OS in mainland China. The character set used in CCDOS was an extended version of GB 1988-80 (G0) and GB/T 2312-1980 (G1), which the basic bitmap font was HZK16 for double byte part at that time. The encoding method for G0 and G1 in CCDOS is the same as that of EUC. In Row 2, CCDOS added 16 segmented hexadecimal digits to the initial 16 empty slots (Cell 1 ~ 16). Other symbols were also added to Row 2, but there is no need to encode them to UCS/Unicode anymore (Cell 67, 68, 79, 80, 93, 94).

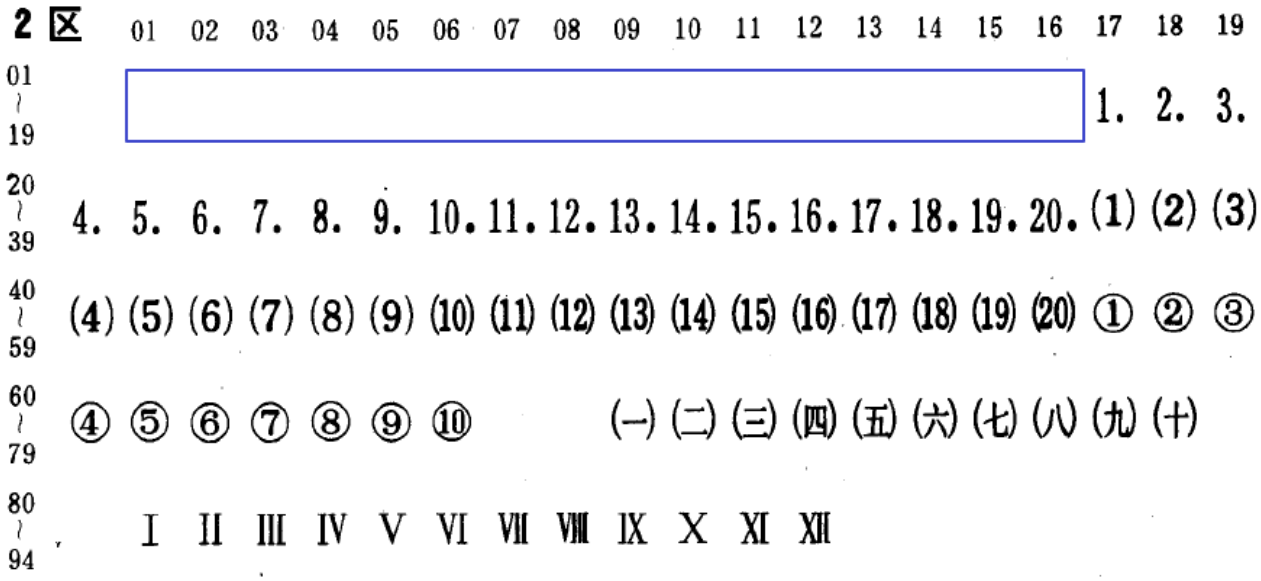

Fig. 0.1 Row 2 of GB/T 2312-1980

Fig. 0.2 Row 2 in CCDOS character set

In GBK, these 16 code points were mapped to ten small Roman numerals and PUAs.

A2	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
A		i 2170	ii 2171	iii 2172	iv 2173	v 2174	vi 2175	vii 2176	viii 2177	ix 2178	x 2179	E586	E587	E588	E589	E58A
B	E58B	1. 2488	2. 2489	3. 248A	4. 248B	5. 248C	6. 248D	7. 248E	8. 248F	9. 2490	10. 2491	11. 2492	12. 2493	13. 2494	14. 2495	15. 2496
C	16. 2497	17. 2498	18. 2499	19. 249A	20. 249B	(1) 2474	(2) 2475	(3) 2476	(4) 2477	(5) 2478	(6) 2479	(7) 247A	(8) 247B	(9) 247C	(10) 247D	(11) 247E
D	(12) 247F	(13) 2480	(14) 2481	(15) 2482	(16) 2483	(17) 2484	(18) 2485	(19) 2486	(20) 2487	① 2460	② 2461	③ 2462	④ 2463	⑤ 2464	⑥ 2465	⑦ 2466
E	⑧ 2467	⑨ 2468	⑩ 2469	E58C	E58D	(一) 3220	(二) 3221	(三) 3222	(四) 3223	(五) 3224	(六) 3225	(七) 3226	(八) 3227	(九) 3228	(十) 3229	E58E
F	E58F	Ⅰ 2160	Ⅱ 2161	Ⅲ 2162	Ⅳ 2163	Ⅴ 2164	Ⅵ 2165	Ⅶ 2166	Ⅷ 2167	Ⅸ 2168	Ⅹ 2169	Ⅺ 216A	Ⅻ 216B	E590	E591	

Fig. 0.3 0xA2XX in GBK

GB 18030-2000 and GB 18030-2005 inherited the mappings in GBK but changed the PUA mappings. GB 18030 also moved U+20AC (€) Euro sign to 0xA2E3 (Cell 67), which mapped to 0x80 in MS-CP936.

A2	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
A		i 2170	ii 2171	iii 2172	iv 2173	v 2174	vi 2175	vii 2176	viii 2177	ix 2178	x 2179	E766	E767	E768	E769	E76A
B	E76B	1. 2488	2. 2489	3. 248A	4. 248B	5. 248C	6. 248D	7. 248E	8. 248F	9. 2490	10. 2491	11. 2492	12. 2493	13. 2494	14. 2495	15. 2496
C	16. 2497	17. 2498	18. 2499	19. 249A	20. 249B	(1) 2474	(2) 2475	(3) 2476	(4) 2477	(5) 2478	(6) 2479	(7) 247A	(8) 247B	(9) 247C	(10) 247D	(11) 247E
D	(12) 247F	(13) 2480	(14) 2481	(15) 2482	(16) 2483	(17) 2484	(18) 2485	(19) 2486	(20) 2487	① 2460	② 2461	③ 2462	④ 2463	⑤ 2464	⑥ 2465	⑦ 2466
E	⑧ 2467	⑨ 2468	⑩ 2469	€ 20AC	E76D	(一) 3220	(二) 3221	(三) 3222	(四) 3223	(五) 3224	(六) 3225	(七) 3226	(八) 3227	(九) 3228	(十) 3229	E76E
F	E76F	Ⅰ 2160	Ⅱ 2161	Ⅲ 2162	Ⅳ 2163	Ⅴ 2164	Ⅵ 2165	Ⅶ 2166	Ⅷ 2167	Ⅸ 2168	Ⅹ 2169	Ⅺ 216A	Ⅻ 216B	E770	E771	

Fig. 0.4 0xA2XX in GB 18030-2000

A2	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
A		ⅰ 2170	ⅱ 2171	ⅲ 2172	ⅳ 2173	ⅴ 2174	ⅵ 2175	ⅶ 2176	ⅷ 2177	ⅸ 2178	ⅹ 2179	E766	E767	E768	E769	E76A
B	E76B	1. 2488	2. 2489	3. 248A	4. 248B	5. 248C	6. 248D	7. 248E	8. 248F	9. 2490	10. 2491	11. 2492	12. 2493	13. 2494	14. 2495	15. 2496
C	16. 2497	17. 2498	18. 2499	19. 249A	20. 249B	(1) 2474	(2) 2475	(3) 2476	(4) 2477	(5) 2478	(6) 2479	(7) 247A	(8) 247B	(9) 247C	(10) 247D	(11) 247E
D	(12) 247F	(13) 2480	(14) 2481	(15) 2482	(16) 2483	(17) 2484	(18) 2485	(19) 2486	(20) 2487	① 2460	② 2461	③ 2462	④ 2463	⑤ 2464	⑥ 2465	⑦ 2466
E	⑧ 2467	⑨ 2468	⑩ 2469	€ 20AC	E76D	〇 3220	〡 3221	〢 3222	〣 3223	〤 3224	〥 3225	〦 3226	〧 3227	〨 3228	〩 3229	E76E
F	E76F	Ⅰ 2160	Ⅱ 2161	Ⅲ 2162	Ⅳ 2163	Ⅴ 2164	Ⅵ 2165	Ⅶ 2166	Ⅷ 2167	Ⅸ 2168	Ⅹ 2169	Ⅺ 216A	Ⅻ 216B	E770	E771	

Fig. 0.5 0xA2XX in GB 18030-2005

The segmented hexadecimal digits are also needed for the scientific calculator, and they are encoded separately from the common Western digits and Latin letters in CCDOS. The full-width Western digits and Latin letters were inherited from GB/T 2312-1980 to put them in Row 3. Therefore, it's better to keep them separately for the legacy mappings in UCS/Unicode.

3 区	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19
01 ~ 19		！	＂	＃	￥	％	＆	＇	（	）	＊	＋	，	－	．	／	０	１	２	３
20 ~ 39	４	５	６	７	８	９	：	；	＜	＝	＞	？	＠	Ａ	Ｂ	Ｃ	Ｄ	Ｅ	Ｆ	Ｇ
40 ~ 59	Ｈ	Ｉ	Ｊ	Ｋ	Ｌ	Ｍ	Ｎ	Ｏ	Ｐ	Ｑ	Ｒ	Ｓ	Ｔ	Ｕ	Ｖ	Ｗ	Ｘ	Ｙ	Ｚ	［
60 ~ 79	＼	］	＾	＿	｀	ａ	ｂ	ｃ	ｄ	ｅ	ｆ	ｇ	ｈ	ｉ	ｊ	ｋ	ｌ	ｍ	ｎ	ｏ
80 ~ 94	ｐ	ｑ	ｒ	ｓ	ｔ	ｕ	ｖ	ｗ	ｘ	ｙ	ｚ	｛	｜	｝	￣					

Fig. 0.6 Row 3 in GB/T 2312-1980

！ ＂ ＃ ￥ ％ ＆ ＇ （ ） ＊ ＋ ， － ． ／ ０ １ ２ ３ ４ ５ ６ ７ ８ ９ ： ； ＜ ＝ ＞ ？ ＠ Ａ Ｂ Ｃ Ｄ Ｅ Ｆ
Ｇ Ｈ Ｉ Ｊ Ｋ Ｌ Ｍ Ｎ Ｏ Ｐ Ｑ Ｒ Ｓ Ｔ Ｕ Ｖ Ｗ Ｘ Ｙ Ｚ ［ ＼ ］ ＾ ＿ ｀ ａ ｂ ｃ ｄ ｅ ｆ ｇ ｈ ｉ ｊ ｋ ｌ
ｍ ｎ ｏ ｐ ｑ ｒ ｓ ｔ ｕ ｖ ｗ ｘ ｙ ｚ ｛ ｜ ｝ ￣

Fig. 0.7 Row 3 in CCDOS Character Set

The Unicode Standard Version 14.0 – Core Specification shows the numeric type of the encoded segmented digits are decimal in p. 850. The segmented hexadecimal digits 0-9 are the same as the decimal ones, so there is no need to encode the segmented hexadecimal digits 0-9 again. There are just six slots after the encoded segmented digits. The proposed new 6 digits are only used for hexadecimal digit representations, and they are treated as numbers in Fig. 2.2 and 2.3, so it's better to use "SEGMENTED HEXADECIMAL DIGIT" as the character names.

1. Proposal

The following is the suggestion to encode these six segmented hexadecimal digits in the *Symbols for Legacy Computing* block. The ordering follows the native hexadecimal digit order.

UCS	Glyph	Character Name	CCDOS Mapping (Row-Cell)
U+1FBFA		SEGMENTED HEXADECIMAL DIGIT A	0xA2AB (02-11)
U+1FBFB		SEGMENTED HEXADECIMAL DIGIT B	0xA2AC (02-12)
U+1FBFC		SEGMENTED HEXADECIMAL DIGIT C	0xA2AD (02-13)
U+1FBFD		SEGMENTED HEXADECIMAL DIGIT D	0xA2AE (02-14)
U+1FBFE		SEGMENTED HEXADECIMAL DIGIT E	0xA2AF (02-15)
U+1FBFF		SEGMENTED HEXADECIMAL DIGIT F	0xA2B0 (02-16)

These hexadecimal digits are only used to represent the specific numerical values, so there is not any case issue. It means we don't need to add anything on these six characters to CaseFolding.txt. For the glyphs, Digits A, C, E and F are related to the capital letters, Digits B and D are related to small letters. Fig. 2.2 and 2.3 shows the relationship clearly. The glyph design shown on the above table is based on *CCDOS Character Set* and Fig. 2.1 cited from Maxim Integrated AN 3210.

The information in NamesList.txt is shown as below.

```
1FBFA   SEGMENTED HEXADECIMAL DIGIT A
    # <font> 0041 latin capital letter a
1FBFB   SEGMENTED HEXADECIMAL DIGIT B
    # <font> 0062 latin small letter b
1FBFC   SEGMENTED HEXADECIMAL DIGIT C
    # <font> 0043 latin capital letter c
1FBFD   SEGMENTED HEXADECIMAL DIGIT D
    # <font> 0064 latin small letter d
1FBFE   SEGMENTED HEXADECIMAL DIGIT E
    # <font> 0045 latin capital letter e
1FBFF   SEGMENTED HEXADECIMAL DIGIT F
    # <font> 0046 latin capital letter f
```

The information in UnicodeData.txt is shown as below and the script values should be "Common" like the encoded segmented digits.

```
1FBFA;SEGMENTED HEXADECIMAL DIGIT A;Nd;0;EN;<font> 0041;10;10;10;N;;;;;
1FBFB;SEGMENTED HEXADECIMAL DIGIT B;Nd;0;EN;<font> 0062;11;11;11;N;;;;;
```

```
1FBFC;SEGMENTED HEXADECIMAL DIGIT C;Nd;0;EN;<font> 0043;12;12;12;N;;;;;
1FBFD;SEGMENTED HEXADECIMAL DIGIT D;Nd;0;EN;<font> 0064;13;13;13;N;;;;;
1FBFE;SEGMENTED HEXADECIMAL DIGIT E;Nd;0;EN;<font> 0045;14;14;14;N;;;;;
1FBFF;SEGMENTED HEXADECIMAL DIGIT F;Nd;0;EN;<font> 0046;15;15;15;N;;;;;
```

The current EastAsianWidth values for the encoded segmented digits are "N", but all the 16 segmented digits are included in the *CCDOS Character Set*, it will be better to use "A" for them like so many Greek and Cyrillic letters. I suggest updating this part in EastAsianWidth.txt as below. If the type designers plan to design them in the font used for Chinese, the glyphs should be designed as full-width.

```
1FBF0..1FBF;A   # Nd    [16] SEGMENTED DIGIT ZERO..SEGMENTED HEXADECIMAL
DIGIT F
```

The line breaking behaviours should be the same as the encoded segmented digits. I suggest updating this part in LineBreak.txt as below.

```
1FBF0..1FBF;NU   # Nd    [16] SEGMENTED DIGIT ZERO..SEGMENTED HEXADECIMAL
DIGIT F
```

The normalization information is shown as below, which should be added to NormalizationText.txt.

```
1FBFA;1FBFA;1FBFA;0041;0041; # (🯊; 🯊; 🯊; A; A; ) SEGMENTED HEXADECIMAL
DIGIT A
1FBFB;1FBFB;1FBFB;0062;0062; # (🯋; 🯋; 🯋; b; b; ) SEGMENTED HEXADECIMAL
DIGIT B
1FBFC;1FBFC;1FBFC;0043;0043; # (🯌; 🯌; 🯌; C; C; ) SEGMENTED HEXADECIMAL
DIGIT C
1FBFD;1FBFD;1FBFD;0064;0064; # (🯍; 🯍; 🯍; d; d; ) SEGMENTED HEXADECIMAL
DIGIT D
1FBFE;1FBFE;1FBFE;0045;0045; # (🯎; 🯎; 🯎; E; E; ) SEGMENTED HEXADECIMAL
DIGIT E
1FBFF;1FBFF;1FBFF;0046;0046; # (🯏; 🯏; 🯏; F; F; ) SEGMENTED HEXADECIMAL
DIGIT F
```

The current VerticalOrientation values for the encoded segmented digits are "R". Despite all the segmented digits are included in *CCDOS Character Set*, there is no need to keep them upright for vertical layouts. At that time, no word processing software could provide the possibility to achieve the vertical layouts in CCDOS. Therefore, the value should be "R" like the encoded segmented digits. I suggest updating this part in VerticalOrientation.txt as below.

```
1FBF0..1FBF   ; R  # Nd    [16] SEGMENTED DIGIT ZERO..SEGMENTED HEXADECIMAL
DIGIT F
```

The proposed characters should be added to SentenceBreakProperty.txt and WordBreakProperty.txt as below.

```
1FBF0..1FBF    ; Numeric  # Nd    [16] SEGMENTED DIGIT ZERO..SEGMENTED
HEXADECIMAL DIGIT F
```

2. Pictures

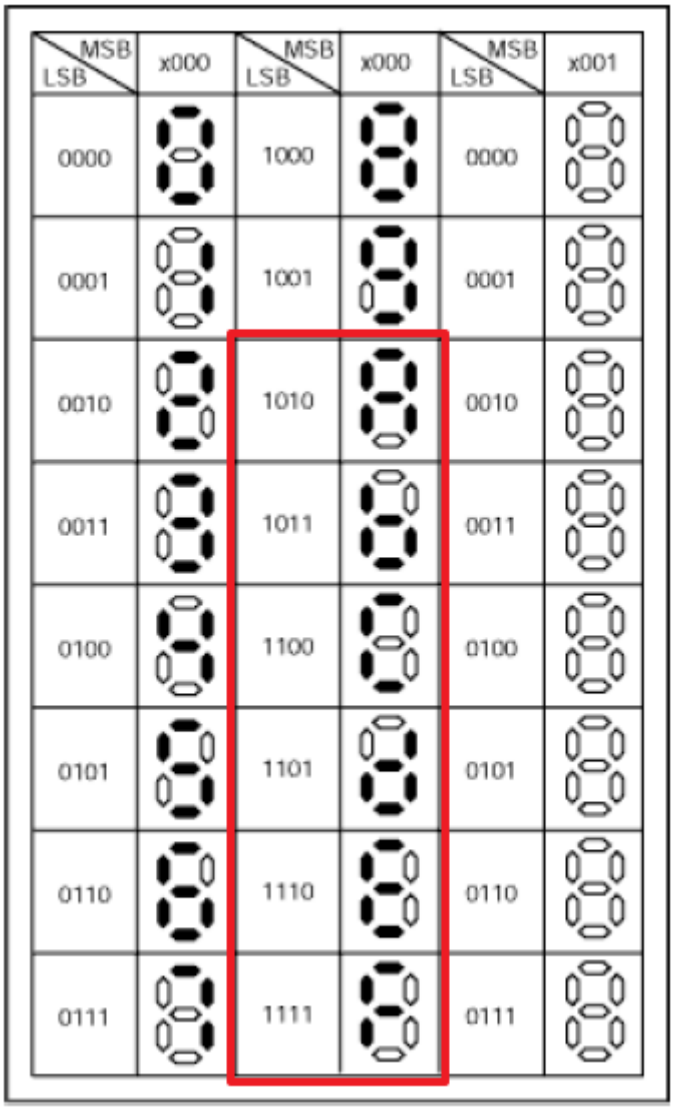

Fig. 2.1 7-segment display font map in [Maxim Integrated Application Note 3210]

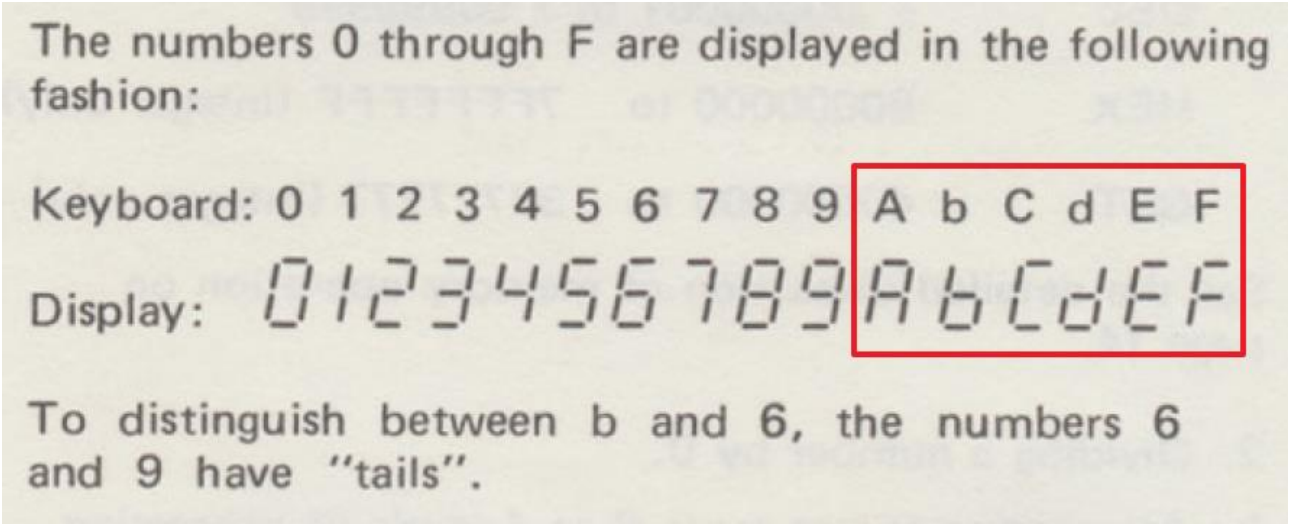

Fig. 2.2 Instruction book of [Texas Instruments electronic calculator TI Programmer 1977] (p. 7)

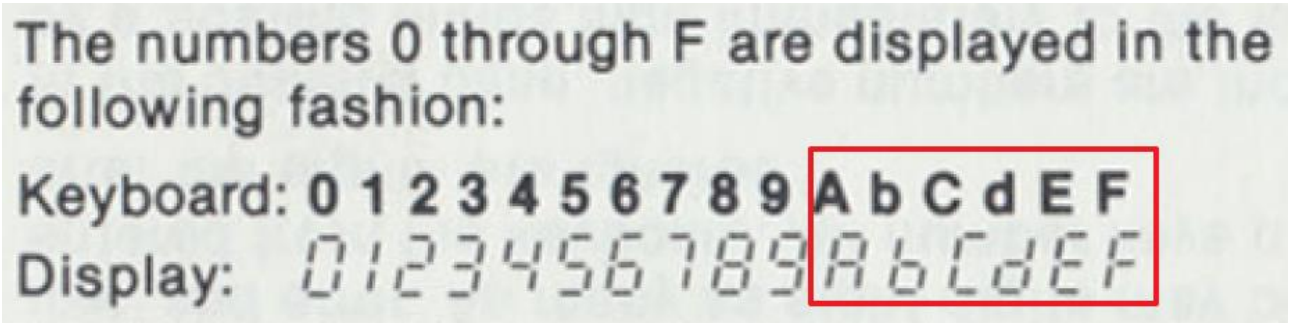

Fig. 2.3 Instruction book of [Texas Instruments electronic calculator TI LCD Programmer 1981](#) (p. 8)

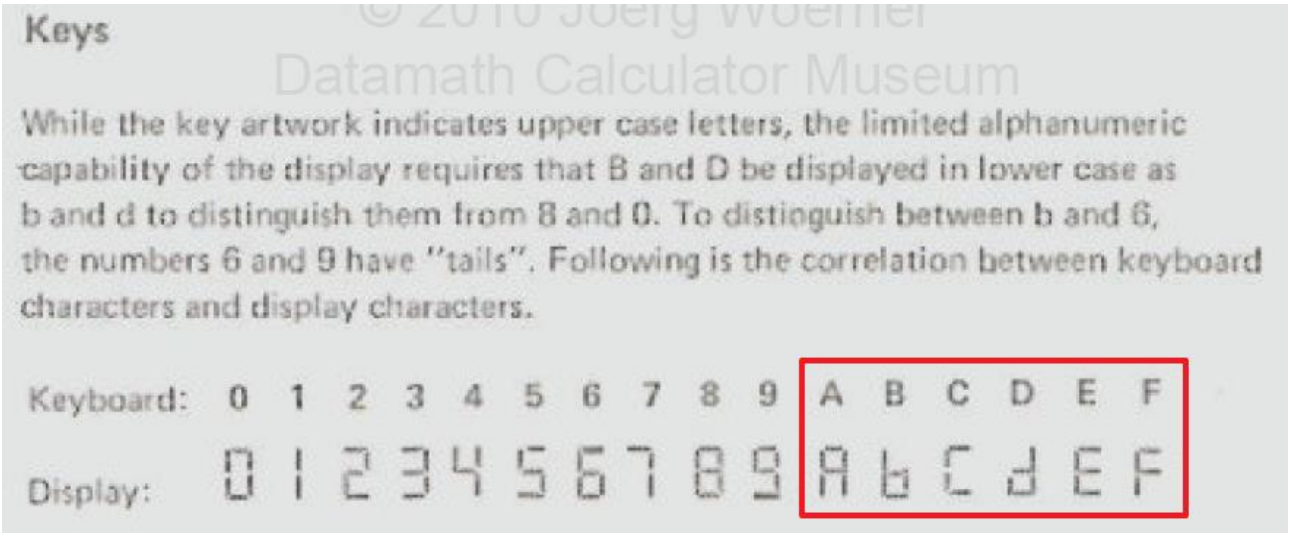

Fig. 2.4 Instruction book of [Texas Instruments hexadecimal calculator/converter SR-22](#) (p. 7)

3. Acknowledgement

Clerk Ma provides the information in CCDOS.
Kushim Jiang provides the instruction books.

(End of Document)

ISO/IEC JTC 1/SC 2/WG 2

PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS

FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646 [1]

Please fill all the sections A, B and C below.

Please read Principles and Procedures Document (P & P) from http://std.dkuug.dk/JTC1/SC2/WG2/docs/principles.html **for guidelines and details before filling this form.**

Please ensure you are using the latest Form from http://std.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html **.**

See also http://std.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html **for latest *Roadmaps*.**

A. Administrative

1. **Title:** ***Proposal to encode segmented hexadecimal digits from Chinese legacy operating system***
2. Requester's name: *Eiso CHAN*
3. Requester type (Member body/Liaison/Individual contribution): *Individual contribution*
4. Submission date: *2022-05-14*
5. Requester's reference (if applicable):
6. Choose one of the following:
   - This is a complete proposal: *YES*
   - (or) More information will be provided later:

B. Technical – General

1. Choose one of the following:
   - a. This proposal is for a new script (set of characters): *NO*
     - Proposed name of script:
   - b. The proposal is for addition of character(s) to an existing block: *YES*
     - Name of the existing block: *Symbols for Legacy Computing*
2. Number of characters in proposal: *6*
3. Proposed category (select one from below - see section 2.2 of P&P document):

A-Contemporary	X	B.1-Specialized (small collection)		B.2-Specialized (large collection)	
C-Major extinct		D-Attested extinct		E-Minor extinct	
F-Archaic Hieroglyphic or Ideographic				G-Obscure or questionable usage symbols	

4. Is a repertoire including character names provided? *YES*
   - a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document? *YES*
   - b. Are the character shapes attached in a legible form suitable for review? *YES*
5. Fonts related:
   - a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard? *Eiso CHAN*
   - b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.): *Eiso CHAN, eisoch@126.com*
6. References:
   - a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? *YES*
   - b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? *YES*
7. Special encoding issues:
   - Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)? *NO*

8. Additional Information:

Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see Unicode Character Database

[1] Form number: N4502-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11, 2005-01, 2005-09, 2005-10, 2007-03, 2008-05, 2009-11, 2011-03, 2012-01)

(http://www.unicode.org/reports/tr44/) and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before? *NO*

   If YES explain

2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)? *YES*

   If YES, with whom? *CCDOS*

   If YES, available relevant documents: *this document*

3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included? *YES*

   Reference: *Contemporary use by specialists and hobbyists.*

4. The context of use for the proposed characters (type of use; common or rare) *Rare*

   Reference:

5. Are the proposed characters in current use by the user community? YES

   If YES, where? Reference: *Worldwide, but particularly in China.*

6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? *NO*

   If YES, is a rationale provided?

   If YES, reference:

7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)? *YES*

8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? *NO*

   If YES, is a rationale for its inclusion provided?

   If YES, reference:

9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters? *NO*

   If YES, is a rationale for its inclusion provided?

   If YES, reference:

10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to, or could be confused with, an existing character? *NO*

    If YES, is a rationale for its inclusion provided?

    If YES, reference:

11. Does the proposal include use of combining characters and/or use of composite sequences? *NO*

    If YES, is a rationale for such use provided?

    If YES, reference:

    Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided? *NO*

    If YES, reference:

12. Does the proposal contain characters with any special properties such as control function or similar semantics? *NO*

    If YES, describe in detail (include attachment if necessary)

13. Does the proposal contain any Ideographic compatibility characters? *NO*

    If YES, are the equivalent corresponding unified ideographic characters identified?

    If YES, reference: