Universal Multiple-Octet Coded Character Set
UCS

Source:	Macao Special Administrative Region, China
Title:	**Submission of Macao's Vertical Extension (UNC Characters), Horizontal Extension, and IVSes Registration for MSCS**
Meeting:	IRG#53 (by circulation)
Status:	Member's submission
Actions required:	To be considered by IRG
Distribution:	IRG Members and Ideographic Experts
Medium:	Electronic
Pages:	9
Appendixes:	4

1   Summary of Request

Macao Special Administrative Region Government (MSARG) is in the process of establishing Macao SAR Information Systems Chinese Character Encoding Scheme (hereinafter referred to as the "Scheme"). The Scheme sets up the exchange framework to define characters used in Macao for information processing and exchange. To address the issue of Macao specific characters, the Scheme includes a Macao Supplementary Character Set, abbreviated as MSCS. MSARG is planning to publish the Scheme as well as MSCS in June 2020 and named the first version of MSCS as MSCS-2020. MSCS-2020 will be used as the information exchange encoding standard among all departments of MSARG first and made publically available for Macao use.

To be aligned with ISO/IEC 10646 for computer processing, MSCS needs to request to IRG/WG2 for (1) vertical extension of 6 characters as UNC submission; (2) horizontal extension of 254 Chinese characters and 16 symbols; and (3) new IVSes registration of 67 variants as well as 64 base characters. Details of these requests will be explained in later sections. The checking of these characters has included all published versions of ISO/IEC 10646 including Extension G (Unicode 13.0 Code Chart) as well as IRG Working Set 2017 v5.0 (IRGN2423). Due to COVID-2019, our proposal cannot be presented in a meeting.

However, due to its urgency, we appreciate IRG experts to provide review and support especially the UNC submissions to be processed quickly.

2 Introduction of the Scheme
2.1 Character Sets and Source References of the Scheme

The Scheme includes the use of three character sets as a collection for information exchange in Macao: (1) the Big-5 character set; (2) Hong Kong Supplementary Character Set (HKSCS) - 2008; (3) Macao Supplementary Character Set (MSCS). The Big-5 character set has been used in Macao since Macao uses the traditional Chinese system. Due to the close connection with Hong Kong SAR, HKSCS characters are also commonly used in Macao and thus should be supported. MSCS will directly use the encoding framework of ISO/IEC 10646.

In the current ISO/IEC 10646 international encoding standard, the source references of all ideographs submitted by MSARG are "MAC-*nnnnn*", in which "nnnnn" is an MSARG-assigned source reference code between 00001 and 99999. Now MSARG establishes a new set of source references for the Scheme. Under the ISO/IEC 10646 international encoding standard, the source references of the Scheme are as follows:

(1) MB-*hhhh* is used to refer to all characters in the Big-5 character set, in which "hhhh" is the hexadecimal Big-5 code. MB0-*hhhh*, MB1-*hhhh*, and MB2-*hhhh* denote symbols, frequently-used ideographs, and less–frequently-used ideographs, respectively, in terms of how they are referenced in ISO/IEC 10646.

(2) MA-*hhhh* is used to refer to all characters already encoded in HKSCS-2008, in which "hhhh" is the corresponding hexadecimal Big-5 code in HKSCS-2008. HKSCS-2008 is the last version of the HKSCS that was published with Big-5 code points.

(3) MC-*nnnnn* is used for characters vertically extended to ISO/IEC 10646, in which "nnnnn" is an MSCS-assigned source reference code between 00001 and 99999, and assigned in sequence.

(4) MD-*hhhh[h]* is used for characters horizontally extended to ISO/IEC 10646, in which "hhhh[h]" is the four- or five-digit hexadecimal code of the character in the ISO/IEC 10646 international standard. For characters in the Basic Multilingual Plane (BMP or Plane 0), four hexadecimal digits are used. For characters in other planes, five hexadecimal digits are used. HKSCS-2016, the latest version of HKSCS, includes 23 ideographs and one symbol horizontally extended to ISO/IEC 10646. Since MSCS also needs to horizontally extend these characters, MDH-*hhhh[h]* is used as the source reference for these characters to differentiate them from other horizontally-extended characters proposed by MSARG.

(5) ME-*hhhh[h]-nnn* is used for character variants with registered IVSes, in which "hhhh[h]" is the four- or five-digit hexadecimal code of the base character in ISO/IEC 10646, and "nnn" is an MSCS-assigned number between 001 and 999. For variants that share the same base character, "nnn" is assigned in sequence.

MSCS includes the following three parts: 1) MSARG's Vertical Extension to ISO/IEC 10646 (source reference: MC-*nnnnn*); 2) MSARG's Horizontal Extension to ISO/IEC 10646 (source reference: MA-*hhhh,* MB-*hhhh*, MD-*hhhh[h]*); and 3) Macao's variants with registered IVSes (source reference: ME-*hhhh[h]-nnn*).

In principle, MSARG only maintains the glyphs and provides fonts for characters in MSCS and does not maintain the glyphs of the characters in Big-5 and HKSCS as a whole. For characters in Big-5 and HKSCS, MSARG will use the fonts in current computer systems. However, for IVSes registration of variants, both variants and the corresponding base characters should be registered. In fact, MSARG also needs to maintain and provide the glyphs of the base characters in Big-5 and HKSCS for IVSes registration. Therefore, MSARG submits the respective base characters in Big-5 and HKSCS in MSCS for IVS registration as horizontal extension to ISO/IEC 10646.

2.2 MSCS-2020

MSARG is finalizing MSCS-2020, which is planned to be published in June 2020 as a part of the Scheme.

MSCS-2020 contains 441 characters, including 425 Chinese characters and 16 symbols. The summary of MSCS-2020 is given as follows:

SN	Current Status	Source Reference	Number		Remarks
			Chinese Character	Symbol	
1	New vertical extension (UNC characters)	MC-*nnnnn*	6		This submission
2	Already vertically extended to ISO/IEC 10646, but the glyphs and source references need to be updated	Change MAC-*nnnnn* to MC-*nnnnn*	86		

SN	Current Status	Source Reference	Number		Remarks
			Chinese Character	Symbol	
3	Horizontal extension of MD (including MDH) characters	MD-*hhhh[h]* MDH-*hhhh[h]*	196	16	This submission
4	Horizontal extension of MA and MB base characters	MA-*hhhh* MB-*hhhh*	58		This submission
5	Already horizontally extended to ISO/IEC 10646, but the glyph and source reference need to be updated	Change MAC-*nnnnn* to MD-*hhhh[h]*	1		
6	New IVSes to be registered, excluding base characters	ME-*hhhh[h]-nnn*	67		This submission
7	IVSes which have been registered in the 2016-08-15 version of the MSARG IVD collection, excluding base characters	ME-*hhhh[h]-nnn*	11		
	Total		**425**	**16**	

The complete set of MSCS-2020 is provided in Appendix A.

This submission includes proposed UNC (SN 1), horizontal extension (SN 3 and 4), and IVSes registration (SN 6). Their data is provided in Appendix B, C and D , respectively.

Since standardization will take time, and MSARG will use the ISO/IEC 10646 coding framework, unencoded MSCS characters in ISO/IEC 10646 will temporarily be assigned with PUA (Private Use Area) code points for font production and exchange before they are encoded in ISO/IEC 10646. Based on the use of PUA code points by HKSCS-2008 and major suppliers of system and software, MSCS will use the PUA code points ranging from EEB8 to F12B for font production for the characters not encoded in ISO/IEC 10646 but accepted by MSCS.

3   Submission of Macao's Characters

3.1   Proposed UNC Characters

In MSCS-2020, 6 characters have not been encoded in ISO/IEC 10646 yet. We checked the latest ISO/IEC 10646 standard (including Ext. G in Unicode 13.0 Code Chart) as well as IRG WS 2017 v5.0 (IRGN2423) and found 2 characters (MC-00137 and MC-00139) are included in IRG WS 2017 v5.0 (IRGN2423). These characters and their source references are listed below. Details of their attributes are listed in Appendix B. MSARG would like for the 6 characters to be encoded as UNC characters. If this is acceptable by IRG experts, the two characters in IRG Working Set 2017 v5.0 (both are proposed by UTC) can be removed from IRG Working Set 2017. But, MSARG is willing to accept other solutions if available.

SN	Source Reference	PUA	Macao's Glyph	Remarks
1	MC-00134	U+EEB8	喹	
2	MC-00135	U+EEB9	譧	It has a similar character: U+8B67 8B67 言 149.13 謙 G3-7429 謙 HB2-F4D4 謙 T2-6A22 謙 K2-6157 "亷" (U+4EB7) and "廉" (U+5EC9) are disunified. In UCS Ideograph Unifiable Component Variations List (IRGN2425), most cases of the two components 337 亷 廉 (SN 337) are disunified. Therefore, this character is proposed as UNC. But other suggestions are welcome.

SN	Source Reference	PUA	Macao's Glyph	Remarks
3	MC-00136	U+EEBA	�	
4	MC-00137	U+EEBB	褢	Included in IRG WS 2017 v5.0 (IRGN2423): SN 02004 (GDM-00085, UTC-02993)
5	MC-00138	U+EEBC	旺	The right component is different from "旺" (U+2512E): 2512E 目 109.4 旺 UCS2003　旺 V2-7D71 In UCS Ideograph Unifiable Component Variations List (IRGN2425), many cases of the components "王" and "壬" (SN 1) are disunified.
6	MC-00139	U+EEBD	塴	Included in IRG WS 2017 v5.0 (IRGN2423): SN 00746 (GDM-00031, UTC-03009)

3.2 Proposed Horizontal Extension

In MSCS-2020, 254 Chinese characters and 16 symbols are to be horizontally extended to ISO/IEC 10646. As they are already coded, we have checked all attributes and supplied our glyph fonts. Details of these characters are provided in Appendix C.

For IVSes registration of variants, both variants and the corresponding base characters should be registered. Therefore, MSARG submits the respective base characters in Big-5 and HKSCS for IVS registration as horizontal extension to ISO/IEC 10646. The basic information of these characters is as follows:

Current Character Set	Source Reference	Number	
		Chinese Character	Symbol
MSCS	MD-*hhhh[h]* MDH-*hhhh[h]*	196	16
Big-5	MB-*hhhh*	16	
HKSCS	MA-*hhhh*	42	
Total		254	16

3.3 Proposed IVSes Registration

In the 2016-08-15 version of the IVD, 11 variants and 10 base characters were registered, along with the MSARG IVD collection itself. The pattern for the sequence identifiers is *M([AB]_[0-9A-F]{4}|C_[0-9]{5}|D_[0-9A-F]{4,5}|E_[0-9A-F]{4,5}_[0-9]{3})*. The format of the sequence identifiers differs slightly from their source reference: 1) The sequence identifiers use underscores in lieu of hyphens per Section 3 of UTS #37; and 2) for the base characters in Big-5, MB is used, and not further distinguished as MB1 or MB2.

MSCS-2020 includes 78 variants, 11 of which were registered the 2016-08-15 version of the MSARG IVD collection. The 67 remaining variants still need to be registered. Because both the variants and their corresponding base characters need to be registered, there are 64 base characters and 67 variants included in this submission. This submission therefore includes 131 proposed new sequences to be added to the registered MSARG IVD collection. All of the variants are included in MSCS-2020. Some base characters are in MSCS-2020 proper, but some are also in Big-5 and HKSCS.

Since the corresponding base characters should also be registered, the number of IVSes in the MSARG collection is as follows:

Status	Number		Description
	Base Character	Variant	
Registered in 2016	10	11	
To be registered	64	67	- They are to be added to the registered MSARG IVD collection. - Pattern for the sequence identifiers is unchanged: *M([AB]_[0-9A-F]{4}|C_[0-9]{5}|D_[0-9A-F]{4,5}|E_[0-9A-F]{4,5}_[0-9]{3})*
Total	74	78	

The information of the base characters in Big-5 (MB) and HKSCS (MA) is summarized as follows:

	Base Characters in Big-5 (MB)	**Base Characters in HKSCS (MA)**
Registered in 2016	1	5
To be registered	15	37
Total	**16**	**42**

Details of the variants to be registered as well as the corresponding base characters are listed in Appendix D.

The temporary web site for IVSes registration is http://www.iso10646hk.net/ivd/MSARG/, which will be changed to another URL in the future.

4   Contact Information

Any queries about the information of the characters, please contact: Mr. Clement Chau (cchau@safp.gov.mo) and Ms. Cheang Pui Pui (ppcheang@safp.gov.mo). We greatly appreciate the effort of all IRG members and ideographic experts.

Appendix A –MSCS-2020 Character Repertoire

Appendix B – List of Proposed UNC Characters

Appendix C – List of Proposed Characters for Horizontal Extension

Appendix D – List of Proposed Characters for IVSes Registration

End of document