

# SUPPORTING CHINESE CHARACTER VARIANTS IN HONG KONG THROUGH IDEOGRAPHIC VARIATION SEQUENCE

Qin Lu<sup>1</sup>, Kwan Hin Cheung<sup>2</sup>, Dan Xiong<sup>1</sup>, Shing Yu<sup>1</sup>, Jian Xu<sup>1</sup>

<sup>1</sup>Department of Computing, <sup>2</sup>Department of CBS

The Hong Kong Polytechnic University

1



THE HONG KONG  
POLYTECHNIC UNIVERSITY

香港理工大學

# OUTLINE

- Introduction
- Background of the Project
- Project Scope and Workflow
- Principles for Specifying Hong Kong Specific Chinese Variants
- Encoding Scheme and Registration
- Progress of the Project
- Conclusion

# INTRODUCTION

- A Chinese character may take different forms due to local preferences: U+9AA8

	Hong Kong Glyph	Mainland China Glyph	Taiwan Glyph
Kai	骨	骨	骨
Song	骨	骨	骨

- Advantages and Limitations of CJK Unification
  - Exchanged text will not change meaning when exchanged
  - Searching and indexing will be easier
  - Different glyphs are supported under different locales
  - Difficult to display different glyphs in the same locale

# STATUS QUO IN HONG KONG

- Hong Kong never had its own independently developed character standard
  - Use of Big5 as the Defacto standard
  - Extended with HKSCS
- HKSARG's endeavour to meet it needs :
  - Published HKSCS in 1999 as extension to Big5
  - Promoted the use of ISO/IEC 10646/Unicode platforms
  - Developed the reference guides on component basis for Chinese computer systems in Hong Kong for both the Song and Kai styles in 2002
  - Component based as there is no other appropriate

𨮒

255	𨮒 0000	4	面緬靨𨮒高..... 9762*7DEC 9766* 9768* 9AD9 0001 0001 0001 0001 1101	(9AD9)與"高"(9AD8)不同。 {見部件 145}
256	止 0000	4	正企步整齒焉歲址..... 6B63* 4F01* 6B65* 6574* 9F52* 7109* 6B72* 5740* 0001 0001 0001 0001 0001 0001 0001 0001	
257	止 0000	4	歧此武疏歸嘴..... 6B67* 6B64* 6B66* 758F* 6B78* 5634* 0001 0001 0001 0001 0001 0001	

# BACKGROUND :

## MORE ON VARIANTS AND UNIFICATION

- Definition of Chinese character variants:
  - The set of different glyph shapes of the same character
  - Do not change the meaning of a character
  - **Example:** ( 雞 鷄 鷄 鸡 ) ( 谿 溪 ) ( 骨 骨 )
- Need for ISO/IEC 10646 unification procedure:
  - Structure analysis, semantics, and glyph variation at different levels
  - Separately coded: ( 雞 鷄 鷄 鸡 ) ( 谿 溪 )
  - Unified: ( 骨 骨 )

# NEW ENABLING TECHNOLOGY

- The Ideographic Variation Sequence (IVS)
  - Define a variant with reference to a defined character
    - Ideograph followed by a variation selector
- Variation selectors coded in ISO 10646: U+E0100 to U+E01EF
  - Total of 240
  - Non-printable characters and cannot be used alone
- IVS: <CJK ideograph, VS>
  - U+8FBB: 辻
  - <U+8FBB, U+E0100>: 辻



# HOW IS IVS USED AND STANDARDIZED

- How to use IVS:
  - The variant selector will be ignored for display if there is no alternative font
  - They can be searched and sorted in the same way as that of the base character
- How to standardize
  - Registration in IVD of Unicode
    - Collection name
    - Permanent site from submitter for public information
    - 3 months public review
- Advantages
  - Suitable for unified characters that cannot be coded otherwise
  - Increased sharing and interoperability cross different platforms than using PUA
- Quite suitable for Hong Kong's Chinese variants
  - Officially define our variant collection against Big5

# PROJECT OBJECTIVE AND SCOPE

## ○ Objectives

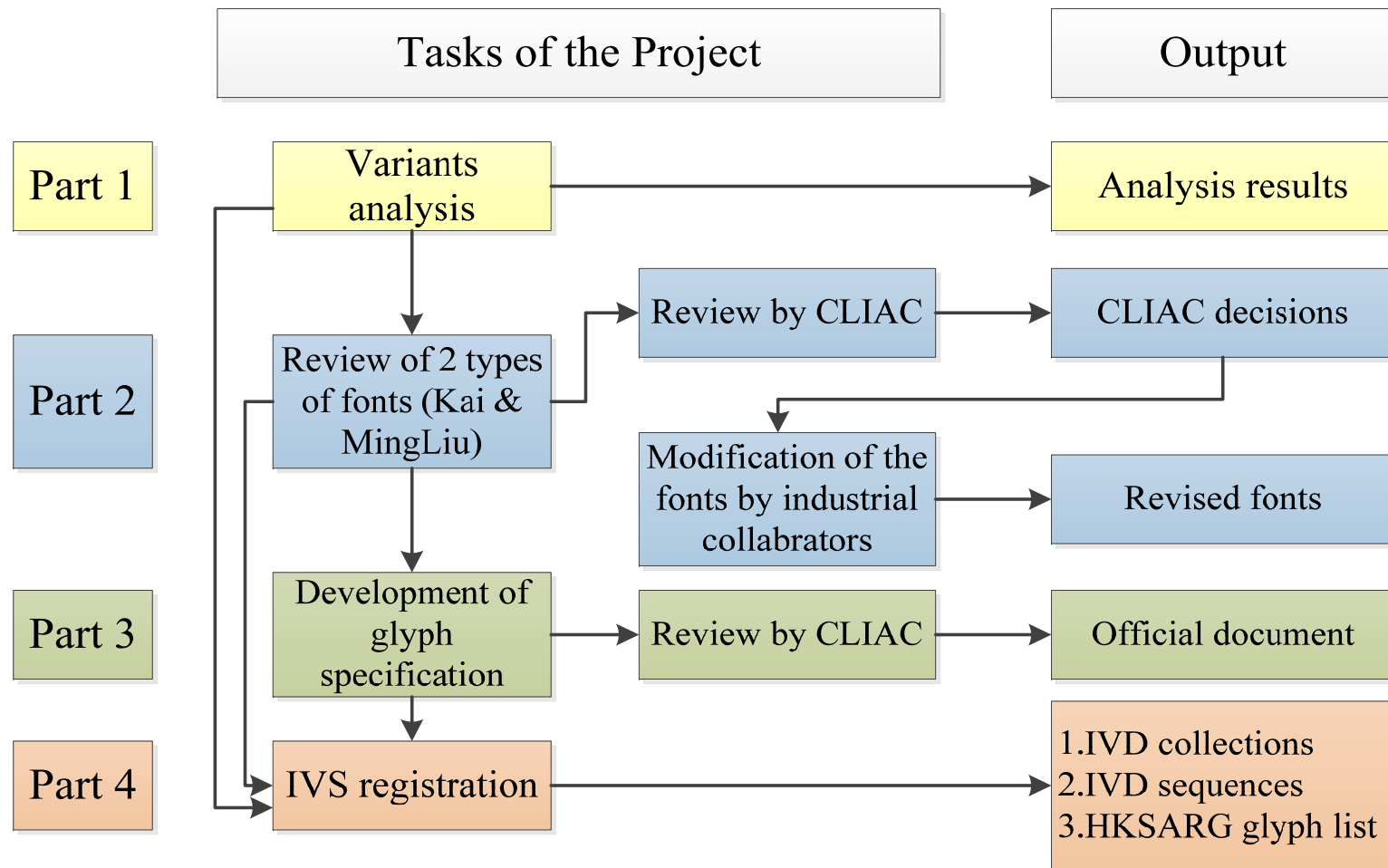
- To specify Hong Kong specific Chinese character variants at the character level
- Under the IVD framework to encode all Chinese character variants under the Big5 coding framework used in Hong Kong so that computer systems can include supports of these variants in different applications

## ○ Project Scope:

- The registration of IVS contains characters in the Big5 repertoire
  - HKSCS characters already adhere to the reference guide of the Hong Kong character glyphs and are not unifiable to any character in Big5



# WORKFLOW : DIAGRAM



# PROJECT COMPONENTS

- Part1: Review and analyze Hong Kong character glyphs with all the characters in the Big5 standard
  - Use different references (by the project team)
- Part2: Produce two types of Chinese character fonts Kai and Song, based on the analysis results (by the project team)
  - *Produce the revised fonts after the variants are identified (by the industrial collaborators)*
  - Review and finalize the deliverables (by CLIAC)
- Part3: Prepare the glyph specification and exhaustive list (by the project team)
  - Review and approve the documents (by CLIAC)
- Part4: Submit the list of the variant characters to the Unicode standard for inclusion in its IVD through Unicode registration
  - Post all required information to the working website of the project (<http://www.iso10646hk.net/ivd/1/> )

# PRINCIPLES TO IDENTIFY VARIANTS

The most important principle is to consider four types of differences:

- 1) Stroke types
- 2) Stroke count
- 3) Relative positions of strokes/components
- 4) Any other difference that may cause confusion of components or word meaning

# DIFFERENCE IN STROKE TYPES

- Basic stroke types : 一 (*Héng, horizontal*), | (*Shù, vertical*), 丿 (*Piě, left slanting*), 丶 (*Diǎn, dot*), ㇇ (*Zhé, turn*)
- Difference in stroke types:

U+	Big5	HK Glyph	Big5 Glyph	Differences
83AB	B2F6	莫	莫	The last stroke is different.
4EA2	A4AE	亢	亢	The first stroke is different.
5316	A4C6	化	化	The 3 <sup>rd</sup> stroke is different.

# DIFFERENCE IN STROKE COUNT

U+	Big5	HK Glyph	Big5 Glyph	Differences
6334	D1C0	梅	梅	The stroke counts of “母” and “母” are different.
6C0F	A4F3	氏	氏	The stroke counts of “丩” and “丩” are different.
82E0	D0A2	苜	苜	The stroke counts of “丩” and “丩” are different.
7B64	DF47	笄	笄	The stroke counts of “丩” and “丩” are different.

# DIFFERENCE IN RELATIVE POSITIONS OF STROKES/COMPONENTS

U+	Big5	HK Glyph	Big5 Glyph	Differences
5317	A55F	北	北	The relative positions of “一” are different.
5E78	A9AF	幸	幸	The relative positions of the horizontal strokes of “𠂇” are different.
5BFA	A678	寺	寺	The relative positions of the horizontal strokes of “土” are different.



## ANY OTHER DIFFERENCE

- Any other difference that may cause confusion of components or word meaning: considered as Hong Kong specific Chinese variant

### 1) Difference in protrusion

U+	Big5	HK Glyph	Big5 Glyph	Differences
5468	A950	周	周	The vertical stroke “   ” of the Hong Kong glyph protrudes downwards.
6025	ABE6	急	急	The horizontal stroke “ 一 ” in the middle of the Hong Kong glyph protrudes rightwards





## ANY OTHER DIFFERENCE

2) Difference in rotation of strokes: the differences make the shapes of the characters quite different

U+	Big5	HK Glyph	Big5 Glyph	Differences
68B2	D5BF	稅	稅	The strokes of “ \ / ” are rotated.
67B0	AC69	枰	枰	The strokes of “ \ / ” are rotated.
706B	A4F5	火	火	The first stroke “ \ ” is rotated.

## ANY OTHER DIFFERENCE

3) Difference in contact of strokes that cause different character semantics

U+	Big5	HK Glyph	Big5 Glyph	Differences
66F0	A4EA			The “一” of the Hong Kong glyph does not touch the right “丨”.
5192	AB5F			The “一” of the upper component of the Hong Kong glyph touches neither the left nor the right “丨”.

# TRIVIAL DIFFERENCES

- Font design differences that may vary from vendor to vendor: classified as trivial differences and will not be considered as variants, eg, “𠃉” and “𠃊”, “𠃋” and “𠃌”

U+	Big5	HK Glyph	Big5 Glyph	Differences
90A3	A8BA	𠃉	𠃊	The first and second strokes of “𠃊” of the Hong Kong glyph touch each other only at the starting point.
827E	A6E3	𠃋	𠃌	The vertical strokes of “𠃌” of the Hong Kong glyph slightly slant inward.

## SCOPE OF THE VARIANT LIST

- The variation list includes all Chinese character glyphs used in Hong Kong which are different from Big5.
- The number of the characters in the list is around 9,000.

## OTHER CONSIDRATIONS

- When Big5 is inconsistent:
  1. When most HK characters containing the component are different from Big5: all related characters are listed in order to specify Hong Kong character glyphs in a consistent way, eg, for 言, Big5 has three glyphs:  
詡 計 訛 (the 3<sup>rd</sup> one is the same as Hong Kong glyph)  
In this case, all characters containing 言 are included in the variant list.
  2. Only a very few characters containing the component are different: eg, 果: In all Big5 characters containing this component, only two have a hook: 罅 轆  
In this case, they are not included in the variant list but listed in another document for reference.



# ENCODING AND REGISTRATION

## ○ Encoding format:

- HKA\_big5-code, eg, 累 : HKA\_B2D6
- “A” refers to Big5 characters used in Hong Kong

UCS	HK Glyph	Big5	Reference Glyph (Song, Ministry of Education of Taiwan)	Description
7D2F	累 HKA_B2D6	累	累	The last stroke of the Hong Kong glyph is different from that of Big5.

## ○ Registration

- All required information will be submitted to Unicode for registration

# PROGRESS OF THE PROJECT

- The variant glyph collection is currently being carried out by the CLIAC.
- The target completion time of the project is **May 2015**.
- It is expected that the Chinese character variant collection for Hong Kong should be included in Unicode before **the end of 2015**.

# CONCLUSION

An ongoing project that applies the IVS and IVD for Hong Kong specific Chinese variant registration:

- To specify Hong Kong specific Chinese character variants at the character level
- To encode all variants under the Big5 coding framework used in Hong Kong so that computer systems can support these variants in different applications