



Two New Chinese Character Standards: HK SCS GB 18030-2000

Dirk Meyer
CJKV Type Development
Adobe Systems Incorporated



18th International Unicode Conference, Hong Kong, April 2001

Overview

- ▶ Similarities ?
- ▶ Hong Kong Supplementary Character Set [HK SCS]: History & Structure
- ▶ Guojia Biao zhun 18030-2000 [GB 18030-2000]: History & Structure
- ▶ HK SCS: Challenges & Implementation
- ▶ GB 18030: Challenges & Implementation
- ▶ Summary / Outlook

Q&A

18th International Unicode Conference Hong Kong, April 2001

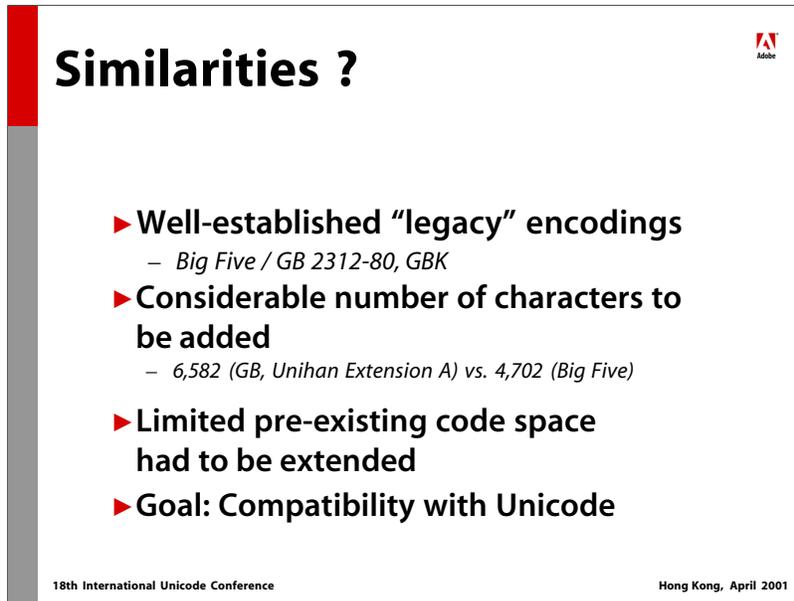
Two New Chinese Character Standards: Hong Kong Supplementary Character Set (HK SCS) / Guojia Biao zhun (Chinese National Standard) 18030-2000 (GB 18030-2000)

The purpose of this presentation is to describe two Chinese “coded character sets” which have been published in August 1999 and in March 2000, respectively.

The two standards’ ultimate goal is to provide an extended set of unique characters to the locale they are meant to be used in. Surprisingly, however, they find similar approaches of how to achieve this. Both have to be based on already existing and well-established existing “legacy” encodings, which limits the number of the code positions at their disposal. At the same time, both are facing the popularity of Unicode as the world’s *lingua franca* for character encoding in the future, and they are trying to maintain or establish compatibility with the universal standard.

The presentation will provide information about the history and the contents of both standards.

To developers who intend to support the two standards and implement their Chinese contents, they offer specific challenges on both encoding and character level. These challenges are closely related to the intended Unicode compatibility of the standards.



Similarities ?

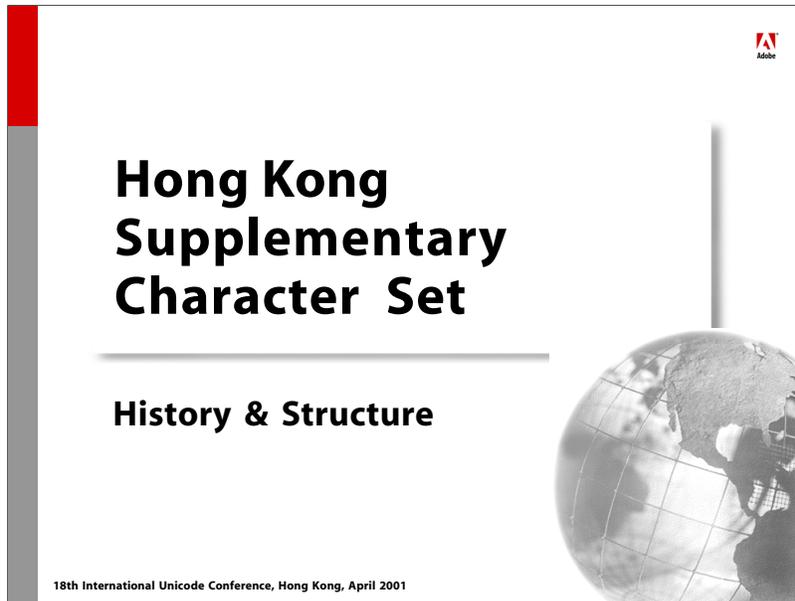
- ▶ **Well-established “legacy” encodings**
 - *Big Five / GB 2312-80, GBK*
- ▶ **Considerable number of characters to be added**
 - *6,582 (GB, Unihan Extension A) vs. 4,702 (Big Five)*
- ▶ **Limited pre-existing code space had to be extended**
- ▶ **Goal: Compatibility with Unicode**

18th International Unicode Conference Hong Kong, April 2001

The question whether there might be any similarities between the two Chinese coded character sets may seem inappropriate, at a first glance. Obviously, the standards provide character coverage for different locales: There is Hong Kong, where “traditional” characters are widely being used, and the People’s Republic of China, where “simplified” Chinese characters were brought into existence. The Cantonese language, spoken in Hong Kong, Southern China and adjacent regions, adds another aspect to this dissimilarity. The integration of the wealth of characters unique to the Cantonese locale should be accommodated in the Hong Kong standard.

However, if we look at the two standards from a different perspective, we find a surprising number of similar obstacles that the two initiatives and their realization were facing:

- Before the standards were introduced well-established “legacy” encodings (Big Five, GB 2312-80/GBK) existed;
- A considerable number of characters was to be added into the new standards (6,582 and 4,702, respectively);
- The code space of pre-existing encodings was limited and had to be extended in order to provide space for the larger character set;
- Both standards acknowledged the importance of the developing Unicode standard and tried to remain compatible or establish compatibility with it.



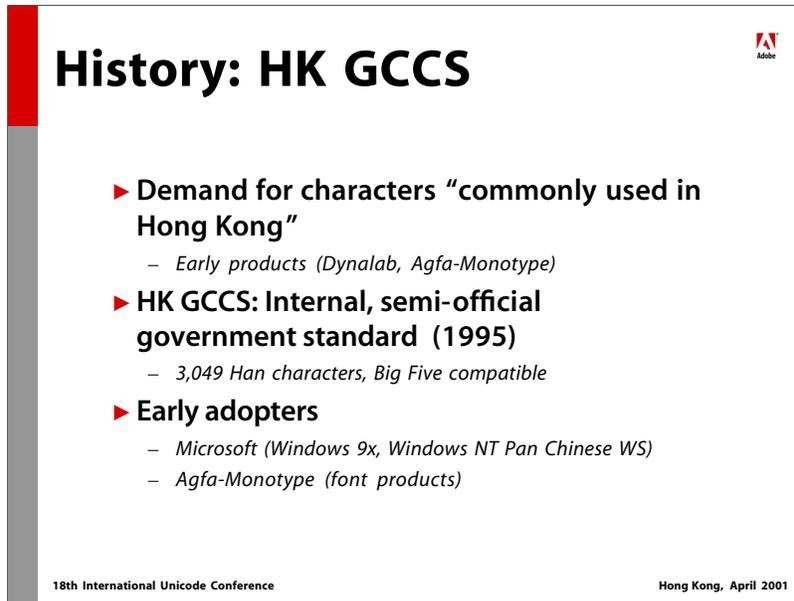
[During the last year's Unicode Conference #16 in Amsterdam, a very detailed description of the history and the contents of the Hong Kong Supplementary Character Set was provided by Qin Lu (Hong Kong Polytechnic University). The slides of this presentation can be viewed at:

<http://www.unicode.org/iuc/iuc16/a16/slides.pdf>]

Historically, computer products lacked acceptable support of “characters commonly used in Hong Kong.” Examples of missing character groups were:

- “Original” Cantonese characters, personal names, and character variants, as they are being used in Hong Kong or the Cantonese-speaking areas of Southern China;
- Han characters of foreign origin (Japan, for example), their use reflecting Hong Kong's special economic and political role in Asia;
- The “simplified” characters of the People's Republic of China.

Neither Big Five nor GB 2312-80 (or its extension “GBK”) were adequately supporting the demands of Cantonese speakers or Hong Kong. Even a full implementation of the Unicode standard would not completely satisfy their needs.



History: HK GCCS

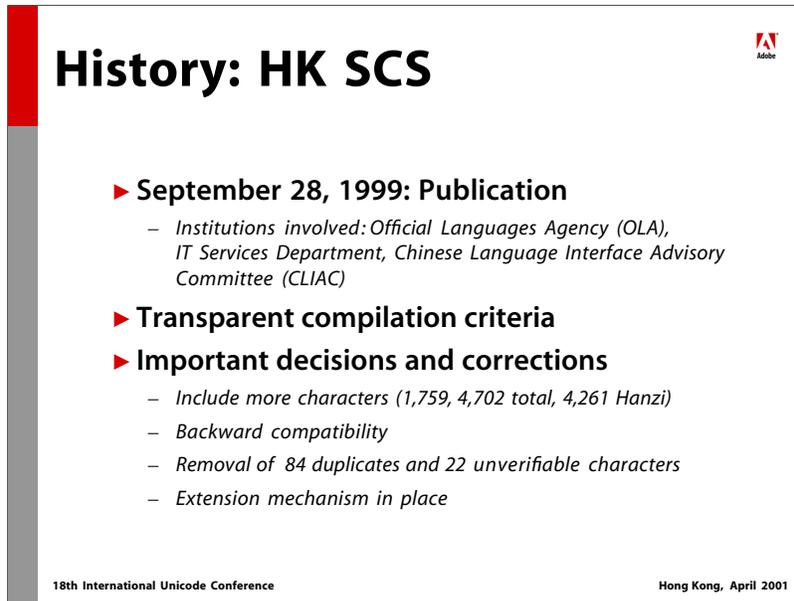
- ▶ **Demand for characters “commonly used in Hong Kong”**
 - *Early products (Dynamlab, Agfa-Monotype)*
- ▶ **HK GCCS: Internal, semi-official government standard (1995)**
 - *3,049 Han characters, Big Five compatible*
- ▶ **Early adopters**
 - *Microsoft (Windows 9x, Windows NT Pan Chinese WS)*
 - *Agfa-Monotype (font products)*

18th International Unicode Conference Hong Kong, April 2001

Roughly ten years ago, type foundries in East Asia started to create products that included characters “frequently or commonly used in Hong Kong.” Due to the lack of early coordination, however, the attempts to support the Hong Kong specific characters remained isolated. Incompatible (font) products reflected different “solutions.” Among those vendor-defined solutions were Dynamlab’s “Hong Kong External Character Collections,” containing either 784 or 644 (1,411) Hanzi (also known as Dynamlab A and Dynamlab B), or Agfa-Monotype’s character sets with 314 and 471 Hanzi.

At the same time, the administration of Hong Kong started to develop the “Hong Kong Government Chinese Character Set” ([HK]GCCS). This “semi-official” specification represented an early attempt to create a common set of Cantonese characters and started to appear in official documents in 1995. Initially, the GCCS was a collection of user-defined characters as they had been in use throughout different branches of the government. In order to prevent incompatibilities and duplications, the government finally decided to adopt the GCCS as a “government-internal” standard. Soon thereafter, support for the GCCS became a feature required for the computers in the Hong Kong government. The GCCS contained a total of 3,049 characters.

In different areas, Microsoft and Agfa-Monotype were two of the early adopters of the GCCS. The former supported the standard (as the “Hong Kong specific End User Defined Characters”) in Chinese versions of Windows 9x and Windows NT 4.0, the latter produced compatible fonts containing even more Cantonese characters than the GCCS itself (3,194).



History: HK SCS

- ▶ **September 28, 1999: Publication**
 - *Institutions involved: Official Languages Agency (OLA), IT Services Department, Chinese Language Interface Advisory Committee (CLIAC)*
- ▶ **Transparent compilation criteria**
- ▶ **Important decisions and corrections**
 - *Include more characters (1,759, 4,702 total, 4,261 Hanzi)*
 - *Backward compatibility*
 - *Removal of 84 duplicates and 22 unverifiable characters*
 - *Extension mechanism in place*

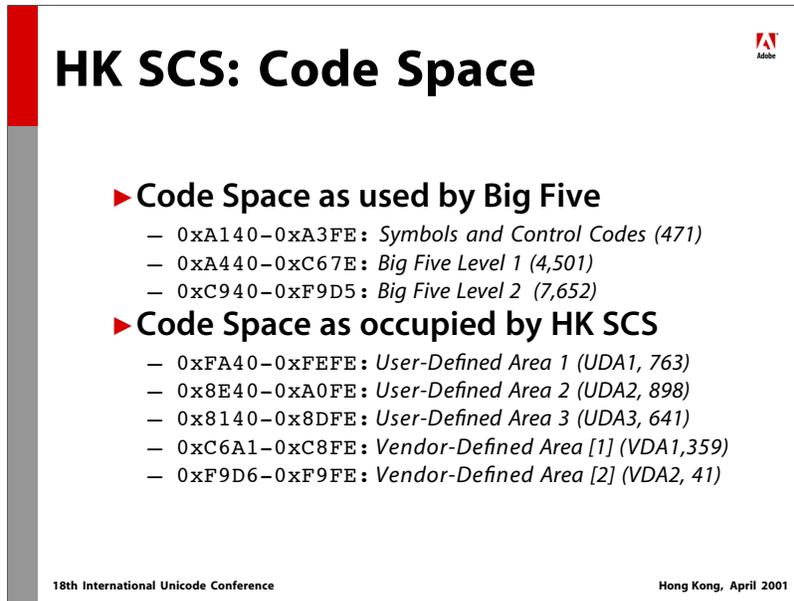
18th International Unicode Conference Hong Kong, April 2001

September 28, 1999, marks the publication date of the “Hong Kong Supplementary Character Set” ([HK] SCS). Preceding steps had been the collection work of the Official Languages Agency (OLA) and the Information Technology Services Department of the government of the Hong Kong Special Administrative Region (SAR), as well as the editorial work of the soon established Chinese Language Interface Advisory Committee (CLIAC).

Interestingly enough, from its early stages on the SCS was discussed and announced on several Web sites. Obviously, nobody underestimated the importance of resolving the conflicts between different or incomplete character standards for the Cantonese locale.

The publication process of the standard, inclusion (and exclusion) criteria for characters, and the contents of the related documents and Web sites left hardly any question unanswered, no matter if related to the characters chosen, to their code points, or to the intentions of the standard’s compilers.

With regard to the GCCS, important correcting steps were taken, like the removal of duplications and “not verifiable” characters. In hindsight, it seems as a wise decision to not have made the GCCS an official standard. Thus, the “supplementary set” left sufficient room to describe, explain and correct shortcomings of the older version. In fact, building on the legacy of the GCCS, the SCS was probably the best possible approach include more required characters, to maintain backward compatibility in order to preserve existing documents, and to extinguish earlier errors.



HK SCS: Code Space

- ▶ **Code Space as used by Big Five**
 - 0xA140–0xA3FE: *Symbols and Control Codes* (471)
 - 0xA440–0xC67E: *Big Five Level 1* (4,501)
 - 0xC940–0xF9D5: *Big Five Level 2* (7,652)
- ▶ **Code Space as occupied by HK SCS**
 - 0xFA40–0xFEFE: *User-Defined Area 1 (UDA1)*, 763
 - 0x8E40–0xA0FE: *User-Defined Area 2 (UDA2)*, 898
 - 0x8140–0x8DFE: *User-Defined Area 3 (UDA3)*, 641
 - 0xC6A1–0xC8FE: *Vendor-Defined Area [1] (VDA1)*, 359
 - 0xF9D6–0xF9FE: *Vendor-Defined Area [2] (VDA2)*, 41

18th International Unicode Conference Hong Kong, April 2001

The SCS contains 4,702 characters, 4,261 of which are Chinese (Hanzi). Non-Han characters also have been added (Cyrillic, Hiragana, Katakana, symbols, numerals, graphic characters, and more).

The original character count of the GCCS of 3,049 was reduced by unifying 84 duplicates and omitting 22 characters “not verifiable.” Added to the reduced number of 2,943 originally contained in the GCCS, were 1,759 new characters, accounting for a total of 4,702 characters of the SCS.

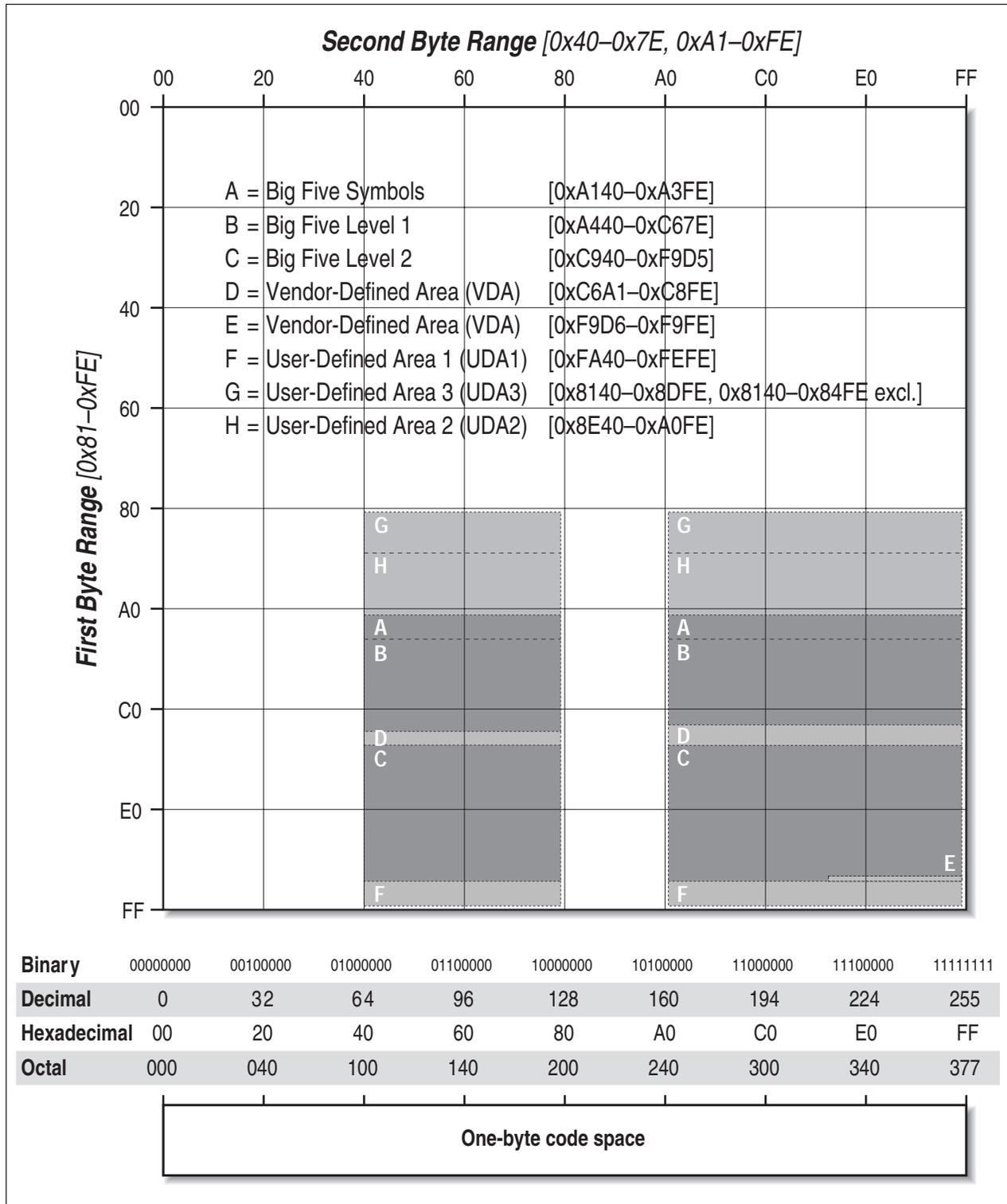
Backward compatibility was maintained in the special cases of unified characters pairs and “not verifiable” characters: The code points formerly occupied by the omitted forms remained reserved, no new character will be assigned to these positions in the future.

The SCS was explicitly designed to fully preserve the code point allocation of Big Five. In addition to that, all 4,702 code points of the SCS can be found within the existing code spaces boundaries of Big Five:

UDA1 (User-Defined Area 1): 0xFA40-0xFEFE for 763 characters,
UDA2 (User-Defined Area 2): 0x8E40-0xA0FE for 2,898 characters,
UDA3 (User-Defined Area 3): 0x8140-0x8DFE for 641 characters,
VDA [1] (Vendor-Defined Area [1]): 0xC6A1-0xC8FE for 359 characters,
VDA [2] (Vendor-Defined Area [2]): 0xF9D6-0xF9FE for 41 characters.

In the vendor-defined areas, the two groups of the so-called “ETen characters” can be found, well-established as a *quasi*-standard and now included in an “official” publication for the first time.

香港增補字符集 [Hong Kong Supplementary Character Set]: Code Space Allocation



HK SCS: Sources

- ▶ <http://www.info.gov.hk/digital21/chi/hkscs/index.html> (Chinese)
- ▶ <http://www.info.gov.hk/digital21/eng/hkscs/index.html> (English)
- ▶ http://www.info.gov.hk/digital21/chi/hkscs/download/new_char_01.pdf (31 additions)
- ▶ http://www.info.gov.hk/digital21/c_index.html (Digital 21, Chinese)
- ▶ http://www.info.gov.hk/digital21/e_index.html (Digital 21, English)

18th International Unicode Conference Hong Kong, April 2001

It is important to mention that there is an extension mechanism in place which is being used to add more characters to the SCS. Under the auspices of the CLIAC, rules for character inclusion and exclusion have been defined, as well as procedures for character submission and review. The publication of newly included characters will happen once a year and will stop after the Unicode Extension B is published. The results of a first inclusion of 31 new characters in April 2000 can be seen at:

http://www.info.gov.hk/digital21/chi/hkscs/download/new_char_01.pdf

In summary: While applying a reasonable encoding strategy and providing numerous characters needed in the Cantonese locale, the SCS maintains, at the same time, code compatibility with Big Five and backward compatibility with its predecessor GCCS.

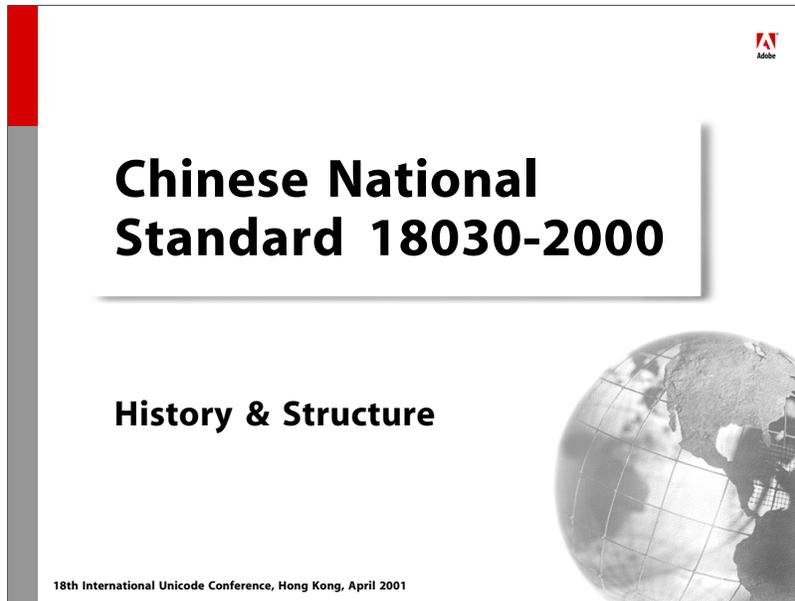
Currently, the characters of the SCS remain to be completely implemented in a future version of Unicode. Present users have to rely on a correct use of Unicode's Private Use Area in order to make full use of this character set.

Documents and mapping data about the SCS, including tables of the “unified” or “not verifiable” characters, and a downloadable font package can be found at:

[http://www.info.gov.hk/digital21/\[chi|eng\]/hkscs/index.html](http://www.info.gov.hk/digital21/[chi|eng]/hkscs/index.html)

“Digital 21 – Information Technology Strategy” is the government of Hong Kong's master plan to make it the “leading digital city in the 21st century.”

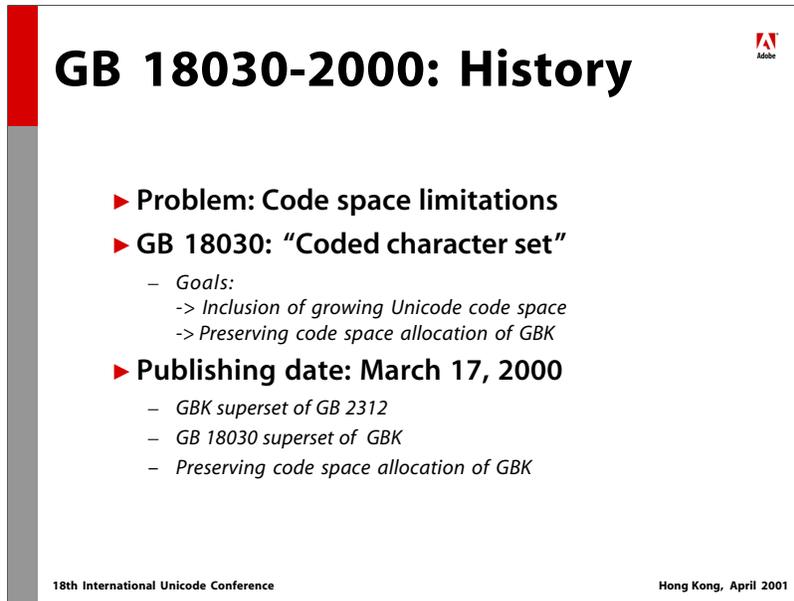
[http://www.info.gov.hk/digital21/\[c|e\]_index.html](http://www.info.gov.hk/digital21/[c|e]_index.html)



Historically, it was the declared interest of the standardization bodies in the People's Republic of China to support the efforts of the ISO/IEC and the Unicode Consortium through publishing a Chinese national standard that was code- and character-compatible with ISO 10646-1/Unicode 2.1. This standard was GB 13000.1-93. Whenever ISO/IEC and the Unicode Consortium implemented changes into their common standard, these changes were subsequently adopted in GB 13000.

In order to remain compatible with GB 2312-80, however, which at the time of GB 18030's publication already was a national standard widely used to represent the Chinese "simplified" characters, the "specification GBK" was created. "GBK" stands for "Guojia biao zhun kuozhan" ["National Standard Expansion"], the official title being "Hanzi neima kuozhan guifan," ["Specifications defining the extensions of internal codes for Chinese ideograms."] GBK is a coded character set that contains a character repertoire very similar to GB 13000, but it uses different encoding areas.

GBK leaves the characters and codes as defined in GB 2312 untouched, characters added later are positioned around it. The additional characters are those of Unicode 2.1's Unified Han portion that go beyond the character repertoire of GB 2312. Thus, code and character compatibility between GBK and GB 2312 were preserved and, at the same time, the complete Unicode Unified Han Character Set becomes available. GBK defines 23,940 code points for 21,886 characters. It also provides mappings to Unicode 2.1. In these mappings, GBK characters not included in Unicode 2.1 as well as "empty" code points are mapped to Unicode's Private Use Area (PUA).



GB 18030-2000: History

- ▶ **Problem: Code space limitations**
- ▶ **GB 18030: “Coded character set”**
 - *Goals:*
 - > *Inclusion of growing Unicode code space*
 - > *Preserving code space allocation of GBK*
- ▶ **Publishing date: March 17, 2000**
 - *GBK superset of GB 2312*
 - *GB 18030 superset of GBK*
 - *Preserving code space allocation of GBK*

18th International Unicode Conference Hong Kong, April 2001

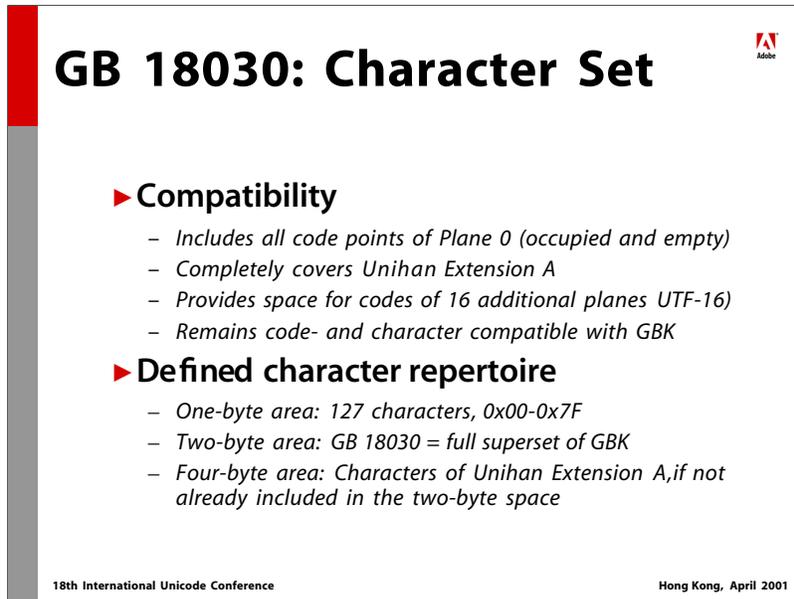
However, considering the packed code space as defined in GBK, it became obvious that there was no space left for a major addition. The 1,894 remaining code points of GBK’s three user-defined areas (UDA 1-3) were not even close to provide sufficient space for the Unihan Extension A, which defines 6,582 additional characters in Unicode 3.0.

A different approach had to be chosen to attempt the complete re-mapping of Unicode 3.0's character repertoire into an extended “legacy” code space of GBK, which now was transformed to become the standard GB 18030-2000.

As its predecessors, GB 18030 is a “coded character set” in that it defines not only a character repertoire, but also standardizes the characters’ code points.

The publishing date of the new “mandatory” Chinese national standard GB 18030-2000 was March 17, 2000. Its official title is “Information technology – Chinese ideograms coded character set for information interchange – Extension for the basic set.”

The main goal of GB 18030 is the harmonious combination of Unicode 3.0’s extended CJK character repertoire (Unihan Extension A) with earlier Chinese national standards or specifications (GB 2312/GBK), while preserving the code space allocation in those.



GB 18030: Character Set

- ▶ **Compatibility**
 - Includes all code points of Plane 0 (occupied and empty)
 - Completely covers Unihan Extension A
 - Provides space for codes of 16 additional planes UTF-16)
 - Remains code- and character compatible with GBK
- ▶ **Defined character repertoire**
 - One-byte area: 127 characters, 0x00-0x7F
 - Two-byte area: GB 18030 = full superset of GBK
 - Four-byte area: Characters of Unihan Extension A, if not already included in the two-byte space

18th International Unicode Conference Hong Kong, April 2001

Using its own language, GB 18030 describes itself as an extension of GB 2312, and as a replacement of the “specification” GBK, version 1.0. The two guiding design principles were that it should remain “encoding standard compatible” with GB 2312, and that it should “completely support all characters of GB 13000 and all characters of CJK Unified Han Extension A.”

In other words, GB 18030 should (1) include Unicode’s Unihan Extension A completely, and (2) provide code space for all occupied code points of Unicode’s plane 0 as well as for the “empty” ones that had not already been included in GBK (same as GB 13000), it should also (3) provide code space for code points of 16 additional planes.

Thus, while being a code- and character compatible “superset” of GBK, GB 18030 also intends to provide space for all remaining Unicode code points. Effectively, it creates a one-to-one relationship between GB 18030 and Unicode’s theoretical encoding space.

In order to accomplish all this, GB 18030 introduces a four-byte encoding mechanism, in addition to already existing one- and two-byte encodings.

GB 18030’s character repertoire includes (1) in the one-byte area: 127 characters as defined on positions 0x00 through 0x7F in GB 11383 (an early version of the standard included the one-byte encoded currency sign “Euro” on position 0x80), (2) in the two-byte area: all characters of GBK’s two-byte area, and (3) in the four-byte area: all characters of the CJK Unified Han Extension A, with the exception of those two-byte encoded characters that have already been included in the two-byte space.

GB 18030: Code Space

▶ **“Legacy”:**

- 0x00 through 0x7F
- 0x81-0xFE (1st byte), 0x40-0x7E and 0x80-0xFE (2nd byte)

▶ **New:**

- 0x8130-0xFE39 (1st/2nd byte),
0x8130-0xFE39 (3rd/4th byte)

0x81308130 – 0x81308139	0x82308130 – 0x82308139
0x81308230 – 0x81308239	...
...	0x8230FE30 – 0x8230FE39
0x8130FE30 – 0x8130FE39	...
0x81318130 – 0x81318139	0xFE308130 – 0xFE30FE39
...	...
0x8131FE30 – 0x8131FE39	0xFE39FE30 – 0xFE39FE39
...	...

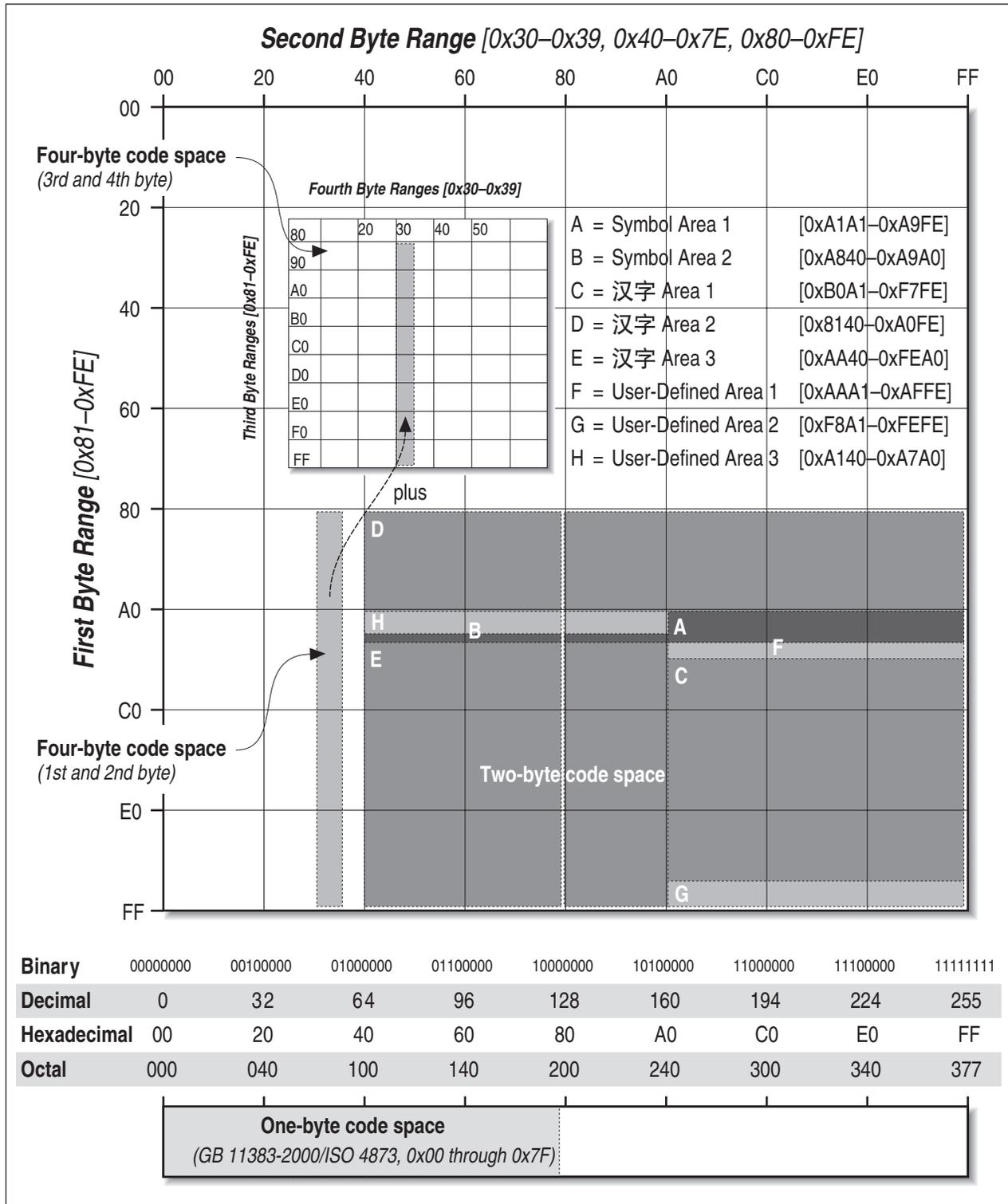
18th International Unicode Conference
Hong Kong, April 2001

Looking at the overall code structure of GB 18030, we can identify known one- and two-byte encoding methods: The one-byte portion applies the coding structure and principles of GB 11383-89 by using 127 code points 0x00 through 0x7F. The two-byte portion includes 23,940 code points by using “first-byte ranges” from 0x81 through 0xFE, and “second-byte ranges” from 0x40 through 0x7E and 0x80 through 0xFE.

As a new feature, the four-byte portion utilizes the code points 0x30 through 0x39 as an additional means to extend the encoding space. This creates an area of four-byte codes, which now contains 1,587,600 code points ranging from 0x81308130 through 0xFE39FE39. The maximum code space, the code mechanism, and the hierarchical sequence of the possible byte combinations are shown above.

In GB 18030’s four-byte mapping area to Unicode, three significant gaps can be identified. (1) The first one reflects that Unicode’s Unihan portion (U+4E00-U+9FA5) is already the major part of GBK/GB 18030. Hence, the last character before and the first one after the Unihan portion, U+4DFF/ U+9FA6, are mapped to neighboring code points of GB 18030. (2) A second mapping gap is related to the Private Use Area (U+E000-U+F8FF). The last Unicode code point before the PUA, U+DFFF, and the first one following those points already used in the two-byte areas of GB 18030, U+E865, map to neighboring GB 18030 code points. All excluded PUA code points can be found in other areas of GB 18030. (3) In the second release of the standard, the surrogate code points (U+D800-U+DFFF) have also been removed from the four-byte mapping area, thus forming the third large gap.

国家标准 [Guojia Biaozhun] 18030-2000: Code Space Allocation





We have shown that the two Chinese “coded character sets” show a number of general similarities in their realization:

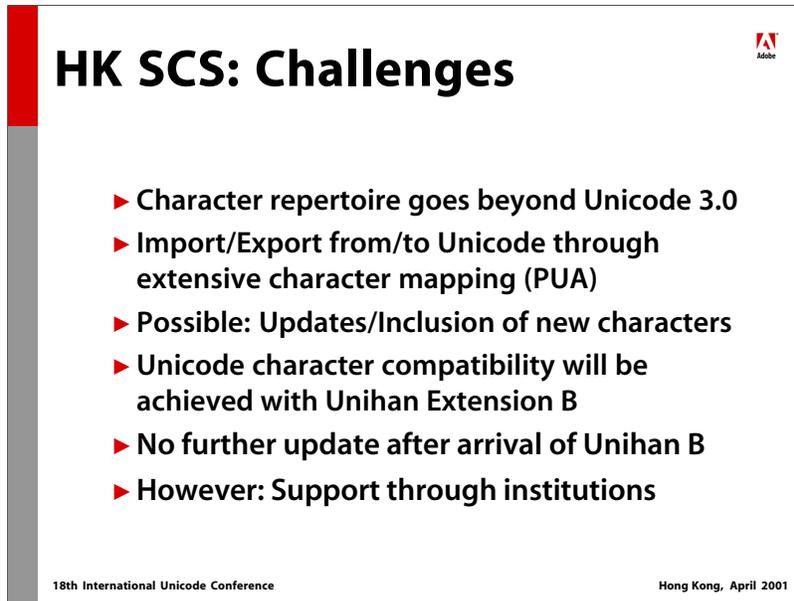
They remain compatible with pre-existing “legacy” standards, i.e. Big Five and GB 2312-80/GBK).

They provide mechanisms to successfully include a large number of additional characters (4,702 and 6,582, respectively).

They also provide rules of how to extend the limited code spaces of their predecessors, which had to be extended in order to provide space for the new and larger character sets.

They both acknowledge the importance of Unicode, either through basing their character repertoire on it (GB 18030) or through providing complete mapping tables to it and encouraging developers to change to Unicode once certain conditions are met, i.e. when the Unihan Extension B has been declared a part of Unicode (SCS).

However, the mechanisms introduced offer challenges of different complexity to developers, which will be described below.



HK SCS: Challenges

- ▶ Character repertoire goes beyond Unicode 3.0
- ▶ Import/Export from/to Unicode through extensive character mapping (PUA)
- ▶ Possible: Updates/Inclusion of new characters
- ▶ Unicode character compatibility will be achieved with Unihan Extension B
- ▶ No further update after arrival of Unihan B
- ▶ However: Support through institutions

18th International Unicode Conference Hong Kong, April 2001

With regard to the implementation of the SCS, product developers in an ever-growing Unicode world are facing some challenges.

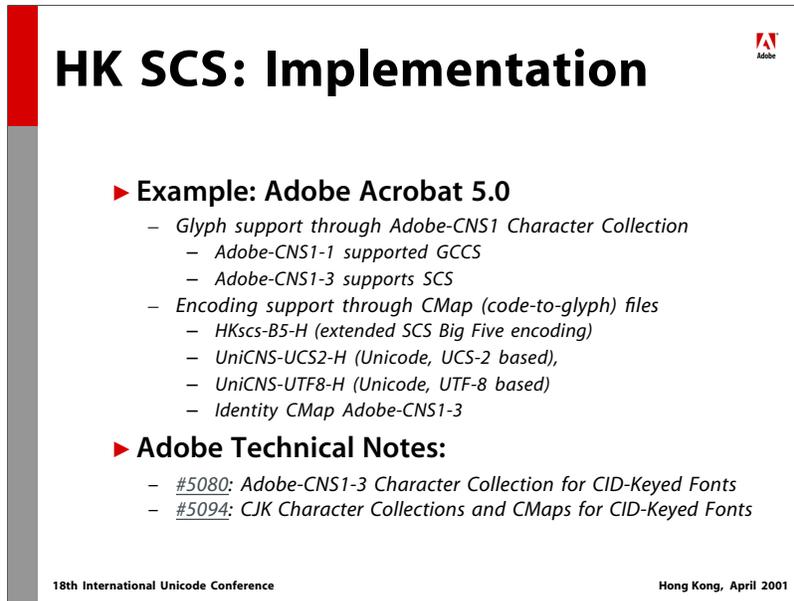
The character repertoire of the SCS contains many new characters that are not yet part of Unicode. This requires additional development in the area of fonts, for example, especially if Unicode – and not the extension of legacy standards – has been the main focus of current development activities.

The fact that a large number of SCS characters do not exist in Unicode, requires the implementation of import and export mechanisms to Unicode through extensive character mapping utilizing Unicode Private Use Area (PUA).

Developers are facing possible updates and the inclusion of additional characters until the arrival of Unihan Extension B. A first addition of 31 characters was implemented in April 2000.

However, these challenges are somewhat limited because the creators of the SCS themselves consider the lifetime of the standard to be finite. With the arrival of Unihan Extension B, character compatibility between Unicode and the SCS will be achieved. After that, no further update of the SCS will take place.

Standardization-related institutions of the Hong Kong government continue to provide a strong encouragement to support only Unicode in future developments.



HK SCS: Implementation

- ▶ **Example: Adobe Acrobat 5.0**
 - *Glyph support through Adobe-CNS1 Character Collection*
 - *Adobe-CNS1-1 supported GCCS*
 - *Adobe-CNS1-3 supports SCS*
 - *Encoding support through CMap (code-to-glyph) files*
 - *HKscs-B5-H (extended SCS Big Five encoding)*
 - *UniCNS-UCS2-H (Unicode, UCS-2 based),*
 - *UniCNS-UTF8-H (Unicode, UTF-8 based)*
 - *Identity CMap Adobe-CNS1-3*
- ▶ **Adobe Technical Notes:**
 - *#5080: Adobe-CNS1-3 Character Collection for CID-Keyed Fonts*
 - *#5094: CJK Character Collections and CMaps for CID-Keyed Fonts*

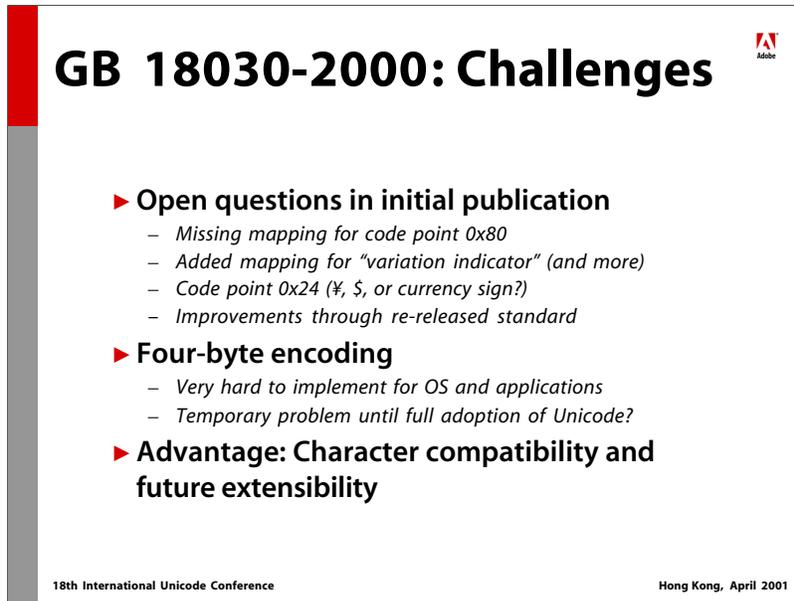
18th International Unicode Conference Hong Kong, April 2001

Adobe Acrobat will start to support the SCS beginning with the release of version 5.0. Internally based on Unicode, Acrobat has to master a significant number of legacy standards. In this process, handling those standards' character codes as well as their glyph repertoire goes hand in hand.

On a glyph level, Adobe's character collection CNS1 started to support Hong Kong specific standard since the arrival of the GCCS. Supplement 1 of Adobe-CNS1 contains 3,309 characters (CIDs 14099 through 17407) that are necessary to provide support for the GCCS and several vendor-specific character sets from Monotype and Dynalab.

Published recently, Supplement 3 of Adobe-CNS1 is adding those characters necessary to fully support the SCS: If Cantonese and other characters have not already been part of earlier Adobe-CNS1 supplements, they are now allocated at CIDs 17606 through 18784. CIDs 18785 through 18845 complete the support for the SCS by allocating positions for Latin or Latin-like characters with diacritic marks and additional symbols. Acrobat 5.0's "substitution fonts" for "traditional" Chinese will be based on character collection Adobe-CNS1-3.

On an encoding level, CMap files have been created that serve both as code-to-glyph files in the context of font technology and as basis for the CMap-like mapping files that are part of Acrobat 5.0. New CMap files include (vertical support versions omitted): HKscs-B5-H (supports the extended Big Five encoding of the SCS), UniCNS-UCS2-H (provides mappings from Unicode [UCS-2] to the Adobe-CNS1-3 glyph repertoire), UniCNS-UTF8-H (same as UniCNS-UCS2-H, but UTF-8-based), and an Identity CMap Adobe-CNS1-3 (provides a continuous mapping to all CNS1-3 glyphs.)



GB 18030-2000: Challenges

- ▶ **Open questions in initial publication**
 - *Missing mapping for code point 0x80*
 - *Added mapping for “variation indicator” (and more)*
 - *Code point 0x24 (¥, \$, or currency sign?)*
 - *Improvements through re-released standard*
- ▶ **Four-byte encoding**
 - *Very hard to implement for OS and applications*
 - *Temporary problem until full adoption of Unicode?*
- ▶ **Advantage: Character compatibility and future extensibility**

18th International Unicode Conference Hong Kong, April 2001

With regard to GB 18030-2000, it will become very important that mapping mechanisms between the new encoding and Unicode are being adopted efficiently. Especially in this area, open issues remained after the initial publication of the standard in April 2000. Among these issues were: The mapping of the added “variation indicator,” the missing mapping for code point 0x80 in the one-byte area, and the question which character to place at code point 0x24 (the Yuan, the Dollar, or the currency sign.) A re-release of Unicode-to-GB18030 mapping data indicates that final decisions with regard to unresolved issues have been made. A new printed version of the standard may very well have been available at the time of this presentation.

The crucial factor for the potential implementation of GB 18030 is to what extent its four-byte encoding can or will be supported through operating systems and applications in the future. Encoding level support for GB 18030’s four-byte approach would result in significant changes affecting current computer environments in numerous and far-reaching ways.

The intended obligatory implementation of the standard by the beginning of the year 2000 was delayed until August 31, 2001. Currently, talks and discussions between experts from all parties involved are under way.

In general, GB 18030 represents the straightforward implementation of a consistent model to establish compatibility between Chinese national standards and Unicode. The advance towards a four-byte encoding mechanism could be anticipated since it became obvious that existing code spaces would become too small for future standard extensions.

GB 18030-2000: Implementation



- ▶ **Again: Acrobat 5.0**
 - *Glyph support through Adobe-GB1 Character Collection*
 - *Adobe-GB1-2 supports GB 13000.1/GBK*
 - *Adobe-GB1-4 supports Unihan A/GB 18030*
 - *Encoding support through CMap (code-to-glyph) files*
 - *GBK2K-H (GB 18030 character set, GB 18030 encoding)*
 - *UniGB-UCS2-H (Unicode, UCS-2 based),*
 - *UniGB-UTF8-H (Unicode, UTF-8 based)*
 - *Identity CMap Adobe-GB1-4*

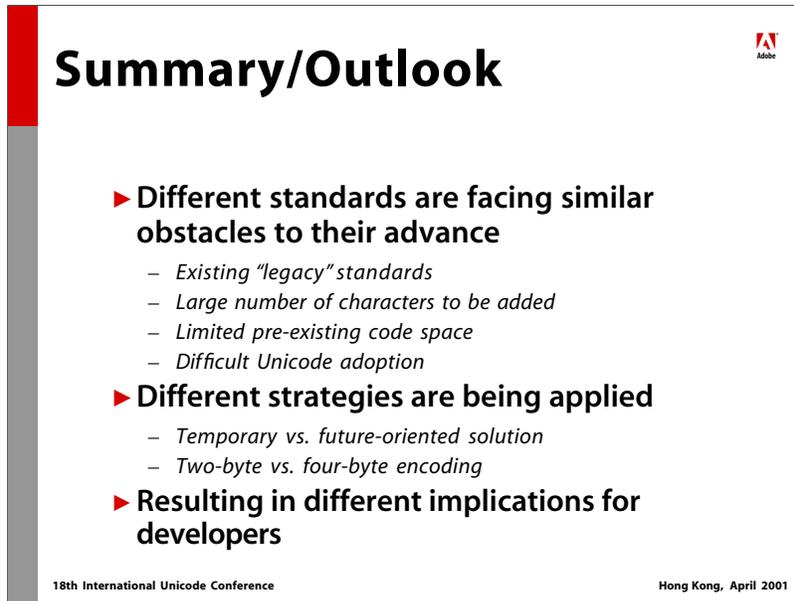
- ▶ **Adobe Technical Notes:**
 - *#5079: Adobe-GB1-4 Character Collection for CID-Keyed Fonts*
 - *#5094: CJK Character Collections and CMaps for CID-Keyed Fonts*

18th International Unicode Conference Hong Kong, April 2001

Again, Acrobat 5.0 may serve as an example for future support of GB 18030. Although there currently is no application support for GB 18030's encoding and character repertoire, Acrobat 5.0 accepts input that utilizes GB 18030's encoding model and provides support for its character repertoire (including the Unihan Extension A).

Supplement 4 of the character collection Adobe-GB1 contains 6,711 additional glyphs at CIDs 22353 through 29063, mainly to provide the characters necessary to support the Unified Han Ideographs Extension A (CIDs 22428 through 29058). Other parts of Adobe-GB1-4 cover additional areas of Unicode 3.0 that are important for the CJK locales (additional Hiragana and Katakana characters and symbols, some of them adjusted for vertical use, the so-called “Hangzhou” or “Suzhou” numerals ten, twenty, and thirty, and the extended Bopomofo glyphs.)

Additional CMap files were added to complete the support for related character standards. Among those are (vertical support versions omitted): GBK2K-H (supports the GB 18030-2000 character set based on a modified GB 18030-2000 encoding), UniGB-UCS2-H (provides mappings from Unicode [UCS-2] to the Adobe-GB1-4 glyph repertoire), UniGB-UTF8-H (same as UniGB-UCS2-H, but UTF-8-based), and an Identity CMap Adobe-GB1-4 (provides a continuous mapping to all GB1-4 glyphs).



Summary/Outlook

- ▶ **Different standards are facing similar obstacles to their advance**
 - Existing “legacy” standards
 - Large number of characters to be added
 - Limited pre-existing code space
 - Difficult Unicode adoption
- ▶ **Different strategies are being applied**
 - Temporary vs. future-oriented solution
 - Two-byte vs. four-byte encoding
- ▶ **Resulting in different implications for developers**

18th International Unicode Conference Hong Kong, April 2001

We have seen that - for the two Chinese character standards - different pre-existing conditions and different intended target locales necessitated different approaches. However, both of them have been facing quite similar obstacles to their completion. Those obstacles included existing “legacy” standards, a large numbers of additional characters, a limited available code space, and a difficult adoption of the current Unicode standard. Quite different strategies were needed to overcome these obstacles.

The creators of the SCS relied on an already existing encoding (Big Five). While preserving existing character allocations, they started to populate areas within the boundaries of the “legacy” standard. The insufficient consistency of an early attempt (GCCS) was accepted and considered during the compilation of the improved standard. Carefully balancing between the current and future needs of the Cantonese locale, the creators of the SCS continue to extend it, but also stress the future importance of Unicode.

The creators of GB 18030 rightfully saw the need for unavoidable, more drastic changes and they decided in favor of a new encoding model that was based on four bytes instead of two. Their premises were to both preserve compatibility between “legacy” and new standards and to include Unicode extensions in GB 18030, thus establishing compatibility in this area, too. At the same time, the new encoding scheme also provided for a complete and final mechanism to include future extensions of Unicode. The standard’s overall significance is not reduced by smaller issues during early stages of its publication. To what extent developers are willing to enter uncharted territory and will provide support for the standard remains to be seen.



Dirk Meyer

dmeyer@adobe.com

Adobe Systems Inc.

Program Manager / CJKV Type Development

345 Park Avenue, M/S W12

San Jose, CA 95110-2704, USA

