中諮文件編號 2002/10

中文界面諮詢委員會

介紹“兩岸四地中文數字化合作論壇”

目的

1. 本文件旨在向各委員介紹“兩岸四地中文數字化合作論壇”會議，及在該會議上就制定中日韓統一表意文字基本子集(“基本子集”)的討論。

背景

2. 「全國信息技術標準化技術委員會」在本年十月三十至三十一日期間在福州安排了一個名為「兩岸四地 IRG 工作研討會」的會議，與會者包括來自中國、香港、澳門和台灣的代表。會議的目的是加強兩岸四地就國際標準化組織轄下的「表意文字小組」工作上的合作，並就彼此在電腦上發展和應用中文編碼標準時遇到共同的問題進行討論。會議並決定將其名稱更改為“兩岸四地中文數字化合作論壇”，其工作範圍載於附件一。會議將會每年輪流在兩岸四地舉行。

3. 在是次福州的會議上，與會代表就制訂基本子集，草擬了一份建議擬稿，並希望兩岸四地在十一月十八至二十二日舉行的「表意文字小組」第二十次會議上，聯名提交該份建議擬稿。

就基本子集的討論

4. 我們曾在「中文界面諮詢委員會」(中諮會) 第十一次會議上，就基本子集的建議作初步討論，並在中諮會轄下的兩個工作小組作深入討論。

5. 「中文資訊科技工作小組」對基本子集的建議表示有所保留。組員認為隨着記憶體技術的發展，小型電子裝置或可於不久的將來便能夠容納整套統一的中日韓表意文字。因此，組員普遍認為沒有需要討論該建議。

6. 「中文電腦用字工作小組」認為，如「表意文字小組」決定接納制訂基本子集的建議，香港須極力爭取將《香港增補字符集》內常用的字符納入該套基本子集內。

7. 香港代表在福州的會議上，已向與會者反映了中諮會轄下的兩個工作小組的意見。基本上，香港對制訂基本子集的建議表示有所保留，而中國、澳門和台灣均表示支持該建議。

福州會議草擬的建議擬稿

8. 我們已就上文第三段提及的建議擬稿的內容，於十一月初諮詢中諮會委員的意見。

9. 總括而言，各委員的意見包括贊成制訂基本子集、不支持制訂基本子集和持保留態度。贊成制訂基本子集建議的理由之一是，並非每個使用表意文字的用戶均須使用全套的 ISO 10646 國際編碼標準。因此，沒有需要強迫用戶使用包含整套 ISO 10646 字集的系統。不支持建議的主要理由是，「表意文字小組」成員之間會因選取適當的字符以納入該基本子集內引起爭論，並會為此耗用不少時間和資源。此外，訂立制訂基本子集的準則亦會相當困難。

10. 歸納中諮會各委員的意見，我們認為當中並無基本或具爭議性的理由，將制訂基本子集的建議變成為完全不值得考慮。另一方面，在福州的會議上所擬備的建議擬稿，是已經考慮過香港方面的意見而草擬的。基於這兩方面的考慮，我們已回覆中國內地表示香港原則上支持該建議擬稿，並根據中諮會委員所提出的意見，對福州會議上草擬的建議擬稿提出相應的修訂(見附件二)。

11. 其後，兩岸四地代表將建議擬稿修改成為附件三的版本，並將其提交給「表意文字小組」第二十次會議作參考。香港代表亦在該次會議上，表達了中諮會各委員對基本子集的意見。最後，「表意文字小組」一致同意制訂基本子集，並草擬了一份制定基本子集的報告，並由小組召集人向管理「表意文字小組」工作的 WG2 提交該報告及尋求 WG2 的進一步指示。報告內就制訂基本子集的建議大綱見附件四。

就制訂基本子集的下一步工作的建議

12. 就決定將哪些香港常用的字符提交「表意文字小組」以納入基本子

集內的工作，我們建議由「中文電腦用字工作小組」負責。有關工作須在二零零三年三月完成。如有需要，我們會就有關工作向中諮會各委員進行諮詢。

13. 歡迎各委員就以上的內容提出任何意見。

資訊科技署
二零零二年十二月

附件一

“兩岸四地中文數字化合作論壇”的工作範圍

Chinese Digitization Forum (CDF)

- CDF 合作的重點是中文編碼及中文數字化應用的共同課題。
- CDF 下設兩個小組
  i. 中文字符集小組（簡稱“字符集組”），負責中文字符集的擴充、整合和子集。在 IRG 活動中，協調兩岸四地，就重大問題保持密切溝通；必要時召開碰頭會。

  ii. 中文數字化促進小組（簡稱“數字化促進組”），負責促進文字層面上的中文信息整理與交流，以推動中文數字化的工作，特別是促進文字工程化的工作。

附件二

Universal Multiple-Octet Coded Character Set UCS

ISO/IEC JTC1/SC2/WG2 IRG N935

Date: 2002-11-16

Source: Title: Status : Distribution: Medium :	China, HKSAR, MSAR, and TCA Proposal on Basic International Ideograph Subset (BIIS) Joint Proposal IRG Members and Ideographic Experts Electronic

Needs

There are over 70,000 CJK unified ideographs encoded in ISO/IEC 10646. Since there are different demands from vendors, implementers and users of CJK unified ideographs, it is necessary to specify a CJK subset containing daily use ideographs for CJK common use. The objectives for producing such a common set are:

a. to lower the cost for users, to provide conveniences to them and meet the day-to-day need;
b. to meet the demands for international information interchange electronically
c. to encourage countries/regions to apply international standards.

We recognize that different applications may need different subsets, a basic subset is needed currently.

Definition

The CJK international basic subset (hereafter abbreviated to Subset) is a coded character set containing basic and most frequently used ideographs from CJK Unified Ideographs and CJK Unified Ideographs Extension A of ISO/IEC 10646.

Acceptance Criteria

a. The repertoire of the Subset should be stable for a long period of time.
b. To speed up the process of defining the Subset, the core of the Subset should be derived from existing basic character sets. Additional characters could then be added to the core on the basis of demonstrated needs, such as frequency of usage, political and cultural

importance.

c. The repertoire of Subset should reflect the need of modern daily use. IRG members are required to provide statistics of ideograph-use frequency, which is based on modern publications, such as newspapers, and elementary education texts, with corpus size over 10,000,000 characters (not limited to ideographs only), while submitting their Subset proposals.
d. Usage frequency of characters should be weighted. Ideographs which cannot be substituted by alphabet-like symbols (such as Japanese kana and Korean hangul) should carry more weight.
e. The repertoires submitted by IRG members should be generated based on, if any, basic ideograph lists issued by their governments or other institutions respectively.
f. Ideographs not in basic ideograph lists but nonetheless needed by IRG members will be considered for inclusion in the Subset. It is suggested that any limit of such ideographs from each IRG member's submission should be discussed and set by IRG.
g. Variant forms of simplified ideographs that are in the basic ideograph lists can be included in the Subset too. Therefore, the simplified and unsimplified variants can be in the Subset if they are in any basic ideograph list such as the *General List of Simplified Hanzi*.
h. Only canonical forms of variants confirmed by their submitters can be contained in the Subset.

It is estimated that the Subset should contain less than 10,000 ideographs according to above criteria.

References: basic ideograph lists

Below are some references of China, HKSAR, MacauSAR and TCA. References from other IRG members are needed.

a. China: *General Purpose Hanzi List of Modern Chinese Ideograms* (National Language Committee, Administration of the Press and Publication)
b. China: *General List of Simplified Hanzi* (National Language Committee, 1986)
c. China: GB 2312-80 Chinese Ideograms Coded Character Set for Information Interchange — Basic Set
d. 香港《常用字字形表》（李學銘主編，香港教育學院 2000 年）
e. Hong Kong: "Hong Kong Supplementary Character Set – 2001" (HKSAR Government 2001)
f. TCA-CNS 11643-1992 1st Plane & 2nd Plane

附件三

Universal Multiple-Octet Coded Character Set UCS

ISO/IEC JTC1/SC2/WG2 IRG N935

Date: 2002-11-16

Source: Title: Status : Distribution: Medium :	China, HKSAR, MSAR, and TCA Proposal on Basic International Ideograph Subset (BIIS) Joint Proposal IRG Members and Ideographic Experts Electronic

Needs

There are over 70,000 CJK unified ideographs encoded in ISO/IEC 10646. Since there are different demands from vendors, implementers and users of CJK unified ideographs, it is necessary to specify a CJK subset containing daily use ideographs for CJK common use. The objectives for producing such a common set are:

a. to lower the cost for users, to provide conveniences to them and meet the day-to-day need;
b. to meet the demands for international information interchange electronically
c. to encourage countries/regions to apply international standards.

We recognize that different applications may need different subsets, a basic subset is needed currently.

Definition

The CJK international basic subset (hereafter abbreviated to Subset) is a coded character set containing basic and most frequently used ideographs from CJK Unified Ideographs and CJK Unified Ideographs Extension A of ISO/IEC 10646.

Acceptance Criteria

a. The repertoire of the Subset should be stable for a long period of time.
b. The repertoire of Subset should reflect the need of modern daily use. IRG members are required to provide statistics of ideograph-use frequency, which is based on modern

publications, such as newspapers, and elementary education texts, with corpus size over 10,000,000 characters (not limited to ideographs only), while submitting their Subset proposals. Ideographs which are represented by phonetic symbols in daily use should carry less weight.

c. The repertoires submitted by IRG members should be generated based on, if any, basic ideograph lists issued by their governments or other authoritative institutions respectively.
d. Ideographs not in basic ideograph lists but nonetheless needed by IRG members will be considered for inclusion in the Subset. It is suggested that such ideographs from each IRG member's submission should be limited to 20 unless strong justifications with high frequency use are provided.
e. Variant forms of simplified ideographs that are in the basic ideograph lists can be included in the Subset too. Therefore, the simplified and unsimplified variants can be in the Subset if they are in any basic ideograph list such as the *General List of Simplified Hanzi*.
f. Only canonical forms of variants confirmed by their submitters can be contained in the Subset.

It is estimated that the Subset should contain less than 10,000 ideographs according to above criteria.

References: basic ideograph lists

Below are some references of China, HKSAR, MacauSAR and TCA. References from other IRG members are needed.

a. China: *General Purpose Hanzi List of Modern Chinese Ideograms* (National Language Committee, Administration of the Press and Publication)
b. China: *General List of Simplified Hanzi* (National Language Committee, 1986)
c. China: GB 2312-80 Chinese Ideograms Coded Character Set for Information Interchange — Basic Set
d. 香港《常用字字形表》（李學銘主編，香港教育學院 2000 年）
e. Hong Kong: "Hong Kong Supplementary Character Set – 2001" (HKSAR Government 2001)
f. TCA-CNS 11643-1992 1st Plane & 2nd Plane

附件四

IRG Report to WG2 on Creation of an ISO 10646 Subset

(Extract of Salient Points)

Considerations for the contents of the common subset

1. The common subset should be geared at providing support for uses such as:

   a. elementary education
   b. unspecialized, general publications such as newspapers, magazines, and novels
   c. textbooks for primary and middle schools
   d. frequently used colloquial (spoken) characters
   e. the most common personal names and place names

Size considerations

2. Among the national and regional standards in current use which provide for the bulk of common use, the relevant character sets are of the following sizes :

China - G0 (level 1)	:	3,755
Taiwan - T1	:	5,412
HKSAR - HKSCS	:	4,818
Japan - J0 (levels 1 and 2)	:	6,356
Korea - K0	:	4,620

3. From government lists for primary and secondary education, there are the following character counts:

China	:	~3,500
Taiwan	:	4,808
HKSAR	:	4,759
Japan	:	1,945
Korea	:	1,800

4. The intersection of the above two sets of counts consists of about 2,500 characters and their union is about 6,300. It is therefore anticipated that a collection of 6,500–7,000 characters would be adequate to cover the needs of the standard subset.

Development Plan

5. Each IRG member is to start with two collections, i.e. Level 1 of its encoded character set and any standard list of characters used in education, preferably

government-issued. The union of the two collections will form the first approximation for the member's submission to IRG. The member will then subtract from and add to this list.

6. Only characters currently encoded in ISO/IEC 10646 may be included. IRG members are strongly discouraged from including characters from Extension B.

7. Justifications for inclusion to the basic list should be available. The justifications need not be detailed. For example, a member may justify characters using the following format:

Place names	:	U+6C39 
People names	:	U+8340 
Colloquial	:	U+4E5C 

8. By March 2003, each member is to send to the IRG a list of the sources it will be using for its list of ideographs, together with an estimate for the size of its final list. The final list must be submitted at least one month before the next IRG meeting so that a merged list can be produced. The IRG will then review the result at the next meeting.