

ISO-IEC JTC1/SC2/WG2

Source : Zhang Zhou Cai, Lu Chin, Lee Kin Hong, Kato Shigenobu,
Ishizaki Shun, Kao Tien-cheu, Ngo Thanh Nhan, Michael Kung,
Tom Naughton - **IRG/Ad Hoc Group for Ideographic
Componsition @Cupertino, CA**
Status : IRG's Contribution for Discussion at WG2
Title : **Ideographic Components and Composition Scheme**

IRG members have concluded that there is a need to find a way to represent some less frequently used ideographs which are not encoded. The following document will try to explain and to justify such a scheme. This scheme consists of two parts: the ideographic components and the composition scheme.

The reasons for such a scheme are :

1. It is used to represent ideographs which are not yet encoded, in conjunction with
 - PU (Private Use)
 - UTF -16
 - UCS - 4
2. It enables composed character representations.
3. It is useful for ideograph look-up
4. It is useful for teaching and analysis of ideographs.
5. It is useful for Character Recognition Systems (OCR, handwritten etc.) and input method designs.

Definitions :

Ideographic Components: are units which can be used to represent ideographs. These components consist of ideographs proper coded in ISO 10646 (BMP) and some basic elements used to form ideographs.

Radicals (IRG definition): are those ideographic components listed in index pages of KXI, DKW, DJW & HYD.

【IRG Rapportuer's Note: They are four typical dictionaries used in Asia. See CJK Annex of ISO 10646.】

Composition Scheme: is a scheme for representing a single ideograph using ideographic components.

Composition Scheme(s):

The composition scheme includes a set of composition functions which can be used recursively to combine ideographic components to generate an ideograph.

The set of composition functions defined are:

1^o: LTR (Left-to-Right)

Function LTR applies to the two following ideographic components to form an ideographic component in such a way that the first component is on the left side and the second component is on the right.

2^o: ATB (Above-to-Below)

Function ATB applies to the two following ideographic components to form an ideographic component in such a way that the second component is placed below the first one.

3^o: AMB (Above-Middle-Below)

Function AMB applies to the three following ideographic components to form an ideographic component in such a way that the first component is on the top, the second component is in the middle and the third component is in the bottom.

4^o: LMR (Left-Middle-Right)

Function LMR applies to the three following ideographic components to form an ideographic component in such a way that the first component is on the left side, the second component is in the middle and the third component is on the right.

5^o: OTI (Outside-to-Inside)

Function OTI applies to the two following ideographic components to form an ideographic component in such a way that the second one is placed inside the first one.

This OTI function can be refined in three possible ways:

- a. **OTI only.**
- b. **OTI can be used in conjunction with the following functions:**
 - OVA - Overall - Around
 - DTE - Down-to-Encompass
 - UTE - Up-to-Encompass
 - RTE - Right-to-Encompass
 - RDE - Right-Down-Encompass
 - RUE - Right-Up-Encompass
 - LDE - Left-Down-Encompass
- c. **The OTI function is replaced by the following functions:**
 - OVA - Overall - Around
 - DTE - Down-to-Encompass
 - UTE - Up-to-Encompass
 - RTE - Right-to-Encompass
 - RDE - Right-Down-Encompass
 - RUE - Right-Up-Encompass
 - LDE - Left-Down-Encompass

Members will carry out further studies to confirm which of these two alternatives should be adopted.

In addition to the above five functions, another function is needed to indicate an ideograph that closely resembles another ideograph but is not exactly the same.

6° : RSM (Resemble)

The function RSM indicates that the following ideograph is not the intended target ideograph but one that resembles it.

The inclusions of this RSM function are subject to further study.

Based on the definitions above, an ideograph can be represented by a "Combined Ideographic Sequence," CIS.

For definition purpose, prefix notations will be used.

$\langle \text{CIS} \rangle ::= \langle \text{symbol2} \rangle \langle \text{CIS}_1 \rangle \langle \text{CIS}_1 \rangle \mid \langle \text{symbol3} \rangle \langle \text{CIS}_1 \rangle \langle \text{CIS}_1 \rangle \langle \text{CIS}_1 \rangle \mid \langle \text{symbol1} \rangle \langle \text{ideographic_components} \rangle$

<CIS₁> ::= <CIS> | <ideographic_component>

<ideographic_component> ::= Coded_Ideograph

<symbol1> ::= RSM

<symbol2> ::= LTR | ATB | OTI

Note: This <symbol2> definition is based on option "a" of function 5 as an example.

<symbol3> ::= LMR | AMB

The definition of CIS is subject to further study.

Ideographic Component Repertoire:

Members have already carried out some feasibility studies and the following is a brief summary of the results:

1. IRG N295 : Korea expressed that such a composition scheme is feasible to express traditional Korean text if two or six characters can be included in the component list.
2. IRG N300 : Vietnam analyzed a sample of 501 characters based on IRG N212 and found that two functions LTR and ATB are sufficient in the description of 87% of the data. Further tests with larger sample sets are in progress and a set of ideographic components for Nom will be presented in the next IRG meeting.
3. IRG N290: TCA had analyzed 10,000 ideographs and found 485 basic components and 557 ideographs proper are sufficient to express these 10,000 ideographs. The same set of basic components and ideographs proper have been able to represent 40,000 ideographs with 98% success.

It is important to point out that a lot of the components used in the studies are already coded. Members will continue their analysis and try to identify the new components, those have not been coded, in order to make the composition scheme workable. For this purpose, two basic tables are required:

1. Radicals coded in ISO/IEC 10646-1
2. Additional Components including the missing radicals (IRG N302 shall be the draft base). These components shall not duplicate either CJK ideographs or vertical extension repertoire.

IRG will submit these tables to WG2.