INTERNATIONAL STANDARD

ISO/IEC 14651

First edition
2001-02-15

AMENDMENT 3
2006-10-15

Information technology — International string ordering and comparison — Method for comparing character strings and description of the common template tailorable ordering

AMENDMENT 3

Technologies de l'information — Classement international et comparaison de chaînes de caractères — Méthode de comparaison de chaînes de caractères et description du modèle commun et adaptable d'ordre de classement

AMENDEMENT 3

Reference number
ISO/IEC 14651:2001/Amd.3:2006(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of the joint technical committee is to prepare International Standards. Draft International Standards adopted by the joint technical committee are circulated to national bodies for voting. Publication as an International Standard requires approval by at least 75 % of the national bodies casting a vote.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

Amendment 3 to ISO/IEC 14651:2001 was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 2, *Coded character sets*.

Information technology — International string ordering and comparison — Method for comparing character strings and description of the common template tailorable ordering

AMENDMENT 3

Page 2, Clause 3

Replace the normative references with the following:

ISO/IEC 10646:2003, *Information technology — Universal Multiple-Octet Coded Character Set (UCS)*

ISO/IEC 10646:2003/Amd.1:2005 *Information technology — Universal Multiple-Octet Coded Character Set — Amendment 1: Glagolitic, Coptic, Georgian and other characters*

Page 17, subclause 6.5

Replace 6.5 with the following:

6.5 Name of the Common Template Table and name declaration

Whenever the Common Template Table is referred externally as a base point in a given context, whether in a process, contract, or procurement requirement, it shall be referenced using the name ISO14651_2005_TABLE1. If another name is used due to practical constraints, a declaration of conformance shall indicate how the correspondence between this other name and the name ISO14651_2005_TABLE1 is taken care of.

The use of a defined name is necessary to manage the different stages of development of this table. This follows from the nature of the reference character repertoire, for which development will be ongoing for a number of years or even decades.

Page 18

Replace Annex A with the following:

Annex A
(normative)

Common Template Table

In order to minimize formatting problems and the risk of errors in reproduction, the common template table is provided separately in a machine-readable file as a normative component of this International Standard. The file name for this language version is different from the normative reference name specified in clause 6.5 of this International Standard due to the existence of file versions commented in other natural languages. The file for this language version can also be retrieved on the ITTF web site at the following URL:

http://www.iso.org/ittf/ISO14651_2005_TABLE1_en.txt

There is an official French version of the file which only differs in its comments (its technical content is identical), and its name is: http://www.iso.org/ittf/ISO14651_2005_TABLE1_fr.txt

NOTE 1 This amendment deprecates, but does not preclude specific reference to, the previous tables, which contained and still contains respectively ordering information From the full repertoire of ISO/IEC 10646-1:2000 and ISO/IEC 10646-2:2001. The previous tables can be found at the following URLs:

[ordering information on the repertoire of characters as defined in ISO/IEC 10646-1:1993 including Amendments 1-9]
http://www.iso.org/ittf/ISO14651_2000_TABLE1.htm

[ordering information on the combined repertoire of characters of ISO/IEC 10646-1:2000 and ISO/IEC 10646-2:2001]
http://www.iso.org/ittf/ISO14651_2002_TABLE1_en.txt

[ordering information on the repertoire of characters as defined in ISO/IEC 10646:2003]
http://www.iso.org/ittf/ISO14651_2003_TABLE1_en.txt

The current Common Template Table reflects the repertoire of characters as defined in ISO/IEC 10646:2003 including its Amendment 1.

NOTE 2 The repertoire targeted by this Amendment 3 to ISO/IEC 14651:2001 is equivalent to the repertoire of *The Unicode Standard Version 4.1, published by The Unicode Consortium*.

When ordering data applicable to other amendments of ISO/IEC 10646:2003 becomes available, this International Standard and specifically its Common Template Table will be amended accordingly to cover the ordering of the additional characters and scripts. To meet cultural requirements of specific communities, delta declarations will have to be applied to the amended table as defined in this International Standard.

ISO_14651_2005_TABLE1 is the name that is used for referring to this table in this version of this International Standard.

Page 24

Include the following new section at the end of Annex B:

B.5. Example 5 – A tailoring for Khmer

The Khmer script is mainly used in Cambodia. The tailoring given below is not included in the CTT (see annex A) itself in order to keep the CTT simple, especially for rare letterforms. E.g. the Khmer ROBAT for which the tailoring below may not be desirable for efficiency reasons, since this letter occurs very rarely, but the tailoring for handling it correctly may affect the efficiency of collation also for texts that do not contain any ROBAT.

```
reorder-after <MAX>


% Khmer:

collating-symbol <S1794_S17C9> % KHMER LETTER BA, KHMER SIGN MUUSIKATOAN

collating-symbol <S1794_S17CA> % KHMER LETTER BA, KHMER SIGN TRIISAP

collating-symbol <S17BB_S17C6> % KHMER VOWEL SIGN U, KHMER SIGN NIKAHIT

collating-symbol <S17B6_S17C6> % KHMER VOWEL SIGN AA, KHMER SIGN NIKAHIT

collating-symbol <C1780>..<C179C>


   % Declaration of Khmer contractions

collating-element <U1794_17C9> from "<U1794><U17C9>" % KHMER LETTER BA, KHMER
SIGN MUUSIKATOAN

collating-element <U1794_17CA> from "<U1794><U17CA>" % KHMER LETTER BA, KHMER
SIGN TRIISAP

collating-element <SW_17CC_1780>..<SW_17CC_17A2> from "<U1780>..<U17A2><U17CC>"

% KHMER LETTER KA, KHMER SIGN ROBAT..KHMER LETTER QA, KHMER SIGN ROBAT

collating-element <SW_17CC_17A5>..<SW_17CC_17B3> from "<U17A5>..<U17B3><U17CC>"

% KHMER INDEPENDENT VOWEL QI, KHMER SIGN ROBAT..KHMER INDEPENDENT VOWEL QAU□,
KHMER SIGN ROBAT

collating-element <U17C6_17BB> from "<U17BB><U17C6>" % KHMER VOWEL SIGN U, KHMER
SIGN NIKAHIT (OM properly spelled)

collating-element <U17BB_17C6> from "<U17C6><U17BB>" % KHMER SIGN NIKAHIT, KHMER
VOWEL SIGN U (OM with the wrong sequence of the characters)

collating-element <U17C6_17B6> from "<U17B6><U17C6>" % KHMER VOWEL SIGN AA, KHMER
SIGN NIKAHIT (AM properly spelled)

collating-element <U17B6_17C6> from "<U17C6><U17B6>" % KHMER SIGN NIKAHIT, KHMER
VOWEL SIGN AA (AM with the wrong sequence of the characters)
```

```
collating-element <U17D2_1780>..<U17D2_179C> from "<U17D2><U1780>..<U179C>"
% COENG, KHMER LETTER KA..COENG, KHMER LETTER QA
collating-element <U17D2_17A5>..<U17D2_17B3> from "<U17D2><U17A5>..<U17B3>"
% COENG, KHMER INDEPENDENT VOWEL QI..COENG, KHMER INDEPENDENT VOWEL QAU

reorder-after <S1794> % KHMER LETTER BA
<S1794_17C9> % KHMER LETTER BA, KHMER SIGN MUUSIKATOAN
<S1794_17CA> % KHMER LETTER BA, KHMER SIGN TRIISAP

reorder-after <S17C5> KHMER VOWEL SIGN AU
<S17BB_17C6> % KHMER VOWEL SIGN U, KHMER SIGN NIKAHIT

reorder-after <S17C6> KHMER SIGN NIKAHIT
<S17B6_17C6> % KHMER VOWEL SIGN AA, KHMER SIGN NIKAHIT

reorder-after <S17D2>
<C1780>..<C1794> % COENG, KHMER LETTER KA..COENG, KHMER LETTER BA
<C1795>..<C179A> % COENG, KHMER LETTER PHA..COENG, KHMER LETTER RO
<C17AB> % COENG, KHMER INDEPENDENT VOWEL RY
<C17AC> % COENG, KHMER INDEPENDENT VOWEL RYY
<C179B> % COENG, KHMER LETTER LO
<C17AD> % COENG, KHMER INDEPENDENT VOWEL LY
<C17AE> % COENG, KHMER INDEPENDENT VOWEL LYY
<C179C>..<C17A2> % COENG, KHMER LETTER VO..COENG, KHMER LETTER QA

reorder-after <SFFFF>
order_start forward;forward;forward;forward

<U1794_17C9> <S1794_17C9>;<BASE>;<MIN>;<U1794_17C9> % KHMER LETTER BA, KHMER SIGN
MUUSIKATOAN
```

```
<U1794_17CA> <S1794_17CA>;<BASE>;<MIN>;<U1794_17CA> % KHMER LETTER BA, KHMER SIGN
TRIISAP

%% The ROBAT contractions should be used only in an "advanced" tailoring for

%% Khmer, since ROBAT is rather rarely used, and these contractions

%% may impact on the efficiency of the key computation even if ROBAT does not

%% occur, since these contractions begin with commonly used letters.

<SW_17CC_1780>..<SW_17CC_17A2>                    "<S179A><S17D2><S1780>..<S17A2>";
"<BASE><VRNT1><BASE><BASE>";"<MIN><MIN><MIN><MIN>";
<SW_17CC_1780>..<SW_17CC_17A2>

% KHMER LETTER KA, KHMER SIGN ROBAT..KHMER LETTER QA, KHMER SIGN ROBAT

<SW_17CC_17A5>..<SW_17CC_17A6>            "<S179A><S17D2><S17A2><S17B7>..<S17B8>";
"<BASE><VRNT1><BASE><BASE><VRNT1><BASE>";"<MIN><MIN><MIN><MIN><MIN><MIN>";
<SW_17CC_17A5>..<SW_17CC_17A6>   %   KHMER   INDEPENDENT   VOWEL   QI,   KHMER   SIGN
ROBAT..KHMER INDEPENDENT VOWEL QII, KHMER SIGN ROBAT

<SW_17CC_17A7>
"<S179A><S17D2><S17A2><S17BB>";"<BASE><VRNT1><BASE><BASE><VRNT1><BASE>";
"<MIN><MIN><MIN><MIN><MIN><MIN>";<SW_17CC_17A7>

% KHMER INDEPENDENT VOWEL QU, KHMER SIGN ROBAT

<SW_17CC_17A8>
"<S179A><S17D2><S17A2><S17BB>";"<BASE><VRNT1><BASE><BASE><VRNT2><BASE>";
"<MIN><MIN><MIN><MIN><MIN><MIN>";<SW_17CC_17A8>

% KHMER INDEPENDENT VOWEL QUK;  KHMER SIGN ROBAT

<SW_17CC_17A9>
"<S179A><S17D2><S17A2><S17BC>";"<BASE><VRNT1><BASE><BASE><VRNT1><BASE>";
"<MIN><MIN><MIN><MIN><MIN><MIN>";<SW_17CC_17A9>

% KHMER INDEPENDENT VOWEL QUU;  KHMER SIGN ROBAT

<SW_17CC_17AA>
"<S179A><S17D2><S17A2><S17BC>";"<BASE><VRNT1><BASE><BASE><VRNT2><BASE>";
"<MIN><MIN><MIN><MIN><MIN><MIN>";<SW_17CC_17AA>

% KHMER INDEPENDENT VOWEL QUUV;  KHMER SIGN ROBAT

<SW_17CC_17AF>..<SW_17CC_17B1>            "<S179A><S17D2><S17A2><S17C2>..<S17C4>";
"<BASE><VRNT1><BASE><BASE><VRNT1><BASE>";"<MIN><MIN><MIN><MIN><MIN><MIN>";
<SW_17CC_17AF>..<SW_17CC_17B1>   %   KHMER   INDEPENDENT   VOWEL   QE,   KHMER   SIGN
ROBAT..KHMER INDEPENDENT VOWEL QOO TYPE ONE, KHMER SIGN ROBAT

<SW_17CC_17B2>
"<S179A><S17D2><S17A2><S17C4>";"<BASE><VRNT1><BASE><BASE><VRNT2><BASE>";
"<MIN><MIN><MIN><MIN><MIN><MIN>";<SW_17CC_17B2>

% KHMER INDEPENDENT VOWEL QOO TYPE TWO;  KHMER SIGN ROBAT
```

```
<SW_17CC_17B3>
"<S179A><S17D2><S17A2><S17C5>";"<BASE><VRNT1><BASE><BASE><VRNT1><BASE>";
"<MIN><MIN><MIN><MIN><MIN><MIN>";<SW_17CC_17B3>

% KHMER INDEPENDENT VOWEL QAU;  KHMER SIGN ROBAT

%%% Khmer OM and AAM (the NIKAHIT should be written after the vowel):

<U17BB_17C6> <S17BB_17C6>;<BASE>;<MIN>;<U17BB_17C6> % KHMER VOWEL SIGN U, KHMER
SIGN NIKAHIT

<U17C6_17BB> <S17BB_17C6>;<BASE>;<MIN>;<U17C6_17BB> % KHMER SIGN NIKAHIT, KHMER
VOWEL SIGN U

<U17B6_17C6> <S17B6_17C6>;<BASE>;<MIN>;<U17B6_17C6> % KHMER VOWEL SIGN AA, KHMER
SIGN NIKAHIT

<U17C6_17B6> <S17B6_17C6>;<BASE>;<MIN>;<U17C6_17B6> % KHMER SIGN NIKAHIT, KHMER
VOWEL SIGN AA

reorder-end
```

Page 25

Replace C.2 with the following:

A.2 Thai string ordering

This annex explains some of the principles behind the tailoring of the CTT given in annex B.5 above, as well as the CTT ordering for Thai (and to some extent Lao).

A.2.1 Thai ordering principles

The widely accepted standard for Thai lexicographical ordering is defined in the Royal Institute Dictionary 2542 B.E. Edition (1999 A.D.), the official standard Thai dictionary. The ordering principles are:

- Words are ordered alphabetically, not phonetically. Consonants order is:

  ก ข ฃ ค ฅ ฆ ง จ ฉ ช ฌ ญ ฎ ฏ ฐ ฑ ฒ ณ ด ต ถ ท

  ธ น บ ป ผ ฝ พ ฟ ภ ม ย ร ฤ ล ฦ ว ศ ษ ส ห ฬ อ ฮ

- Vowels, and nikhahit, are also ordered by written forms, not by sounds. Vowels and nikhahit order is:

  ◌ำ -ะ ◌ั -า -ๅ ◌ิ ◌ี ◌ึ ◌ื ◌ุ ◌ู เ- แ- โ- ใ- ไ- ◌ๆ

  อ ว ย are always ordered as consonants, although they sometimes act as vowels.

  ๅ is a long-legged variant of า, used with the long-legged consonants ฤ and ฦ: ฤๅ and ฦๅ.

  ◌ำ is logically an า followed by a ◌ิ. However, the Unicode compatibility decomposition of the precomposed character is to a ◌ิ followed by an า, so this misspelling must be handled as well.

- No syllable structure or word boundary analysis is required, as Thai lexicons are ordered alphabetically, not phonetically. Note that Thai normally does not use any word separator, except, and exceptionally, zero width space.

- String comparison is performed from left to right, but considering initial consonants before vowels in the same syllable. Leading vowels (เ- แ- โ- ใ- ไ-, corresponding to characters U+0E40-U+0E44), which are written before the consonant, must be considered after the consonant. Therefore, rearrangement (in some way) is needed before comparison.

- Tones and diacritics are ignored at level 1. At level 2 their order is:

  ◌็ ◌์ ◌๋ ◌่ ◌้ ◌๊ ◌๎

- Since Thai, Lao, and Khmer are uncased, it may seem that the third level is not needed for Thai, Lao, and Khmer string ordering. However, the third level is used to differentiate LAKKHANGYAO (ๅ) as a variant of SARA AA (า), since similar variants in other scripts are differentiated in a like manner at level 3, as well as for other variation cases in Thai, Lao, and Khmer.

- When Thai punctuation marks (๏ ๚ ๛ ◎ ๆ) are concerned, another level of weights is required for them. This corresponds to the fourth level in the Common Template Table. In string ordering, punctuation marks are less significant than any tone marks and diacritics, and must be ignored in all the first three levels. Note that PAIYANNOI (ฯ) and THAI CHARACTER MAIYAMOK (repeat mark ๆ) are regarded and ordered as punctuation marks, not letters, despite their Unicode general category as “Lo” and “Lm” respectively. For example, “ข้างๆ, ข้างกบ, ข้างๆ คูๆ, ข้างจัน” is a valid order in the Royal Institute Dictionary. In the first level, the considered weights correspond to ขาง, ขางกบ, ขางคู, and ขางจัน respectively.

- The ten Thai decimal digits (๐ ๑ ๒ ๓ ๔ ๕ ๖ ๗ ๘ ๙), each semantically equivalent to Arabic digit 0-9, respectively. Their weights are then equal to their corresponding Arabic digit in the first level, and are different in the second level, to distinguish script.

A.2.2 Vowel/consonant rearrangement

Regarding the handling of pre-vowels, either a collation preparation or collating-element grouping (as in the tailoring in annex B.5 above) is required. The collation preparation scans the string once and swaps every leading Thai vowel with its succeeding character (ideally only if the succeeding character is a Thai consonant). The prepared string is then passed to the normal weight calculation process. Another way to manage this is by means of collating-element formation – the approach taken both by the CTT of this standard and by the collation weighting table of the Unicode Collation Algorithm (UTS #10). Every possible pair of leading vowel and consonant is defined as a collating-element, whose weight equals to that of the rearranged substring. In addition, since two เ in sequence look just like a แ, two เ in sequence should be handled just like a แ.

Note that the rearrangement of each leading vowel is simply performed with its immediate succeeding consonant. No consonant cluster analysis is needed. Indeed, doing so would result in ambiguities or yield a different order than that specified in the Royal Institute Dictionary. For example:

1. Ambiguities: The problem with ambiguity is illustrated by the word "เพลา". It has two potential pronunciations: either as a two-syllable word, "phe-la" (meaning "time"), or as a one-syllable word, "phlao" (meaning "axle" or "abate"). A rearrangement algorithm which follows the distinct pronunciation of the potential cluster 'พล' in this string would result in distinct keys, "พเลา" and "พลเา", and therefore different weights, which are equally legal. Both words need to have the same weight to be sortable, however.

2. Non-conforming ordering: To illustrate the difference in ordering caused by the treatment of consonant clusters, consider these words, shown in conforming order: "เพล, เพลง, เพศ". The correct rearrangement ignores any clusters and results in the following: "พเล, พเลง, พเศ", which sorts in the order shown. If, however, pairs of consonants that form legal clusters were grouped as single collation elements (regardless of actual pronunciation where the potential pronunciation is ambiguous), then the results of rearrangement would be "<พล>เ, <พล>เง, พเศ", which would yield the (non-conforming) ordering "เพศ, เพล, เพลง". Again, if actual clusters were grouped as single collation elements (with some disambiguation effort), then the results of rearrangement would be "พเล, <พล>เง, พเศ", which would yield the (non-conforming) ordering "เพล, เพศ, เพลง".