NORME INTERNATIONALE

ISO/IEC 14651

Sixième édition
2020-12

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

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

Numéro de référence
ISO/IEC 14651:2020(F)

ISO copyright office
Case postale 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Genève
Tél.: +41 22 749 01 11
E-mail: copyright@iso.org
Web: www.iso.org

Publié en Suisse

Sommaire

Page

Avant-propos ..... **iv**
Introduction ..... **v**
1 Domaine d'application **1**
2 Références normatives **2**
3 Termes et définitions **2**
4 Symboles et conventions **3**
5 Conformité **3**
6 Comparaison de chaînes **4**
6.1 Prétraitement des chaînes de caractères avant comparaison 4
6.2 Construction des clés et comparaison 5
6.2.1 Préliminaires 5
6.2.2 Méthode de référence de construction des clés 6
6.2.3 Méthode de comparaison de référence pour le tri des chaînes de caractères 8
6.2.4 Méthode de comparaison de référence pour le tri des chaînes de caractères 9
6.3 Table-modèle commune: composition et interprétation 10
6.3.1 Généralités 10
6.3.2 Règles de syntaxe BNF pour la table-modèle commune de l'Annexe A 10
6.3.3 Contraintes de forme 12
6.3.4 Interprétation des tables adaptées 14
6.3.5 Évaluation des tables de poids 15
6.3.6 Conditions d'équivalence de tables spécifiques 15
6.3.7 Conditions d'équivalence des résultats 16
6.4 Déclaration d'un delta 16
6.5 Nom de la table-modèle commune et déclaration de nom 18
Annexe A (normative) **Table-modèle commune** **19**
Annexe B (informative) **Exemples de deltas d'adaptation** **21**
Annexe C (informative) **Prétraitement** **31**
Annexe D (informative) **Annexe didactique sur les solutions apportées par le présent document aux problèmes de tri lexical** **47**
Annexe E (informative) **Recherches et correspondances floues** **51**
Bibliographie **53**

Avant-propos

L'ISO (Organisation internationale de normalisation) et l'IEC (Commission électrotechnique internationale) forment le système spécialisé de la normalisation mondiale. Les organismes nationaux membres de l'ISO ou de l'IEC participent au développement de Normes internationales par l'intermédiaire des comités techniques créés par l'organisation concernée afin de s'occuper des domaines particuliers de l'activité technique. Les comités techniques de l'ISO et de l'IEC collaborent dans des domaines d'intérêt commun. D'autres organisations internationales, gouvernementales et non gouvernementales, en liaison avec l'ISO et l'IEC, participent également aux travaux.

Les procédures utilisées pour élaborer le présent document et celles destinées à sa mise à jour sont décrites dans les Directives ISO/IEC, Partie 1. Il convient, en particulier de prendre note des différents critères d'approbation requis pour les différents types de documents ISO. Le présent document a été rédigé conformément aux règles de rédaction données dans les Directives ISO/IEC, Partie 2 (voir www.iso.org/directives).

L'attention est attirée sur le fait que certains des éléments du présent document peuvent faire l'objet de droits de propriété intellectuelle ou de droits analogues. L'ISO et l'IEC ne sauraient être tenues pour responsables de ne pas avoir identifié de tels droits de propriété et averti de leur existence. Les détails concernant les références aux droits de propriété intellectuelle ou autres droits analogues identifiés lors de l'élaboration du document sont indiqués dans l'Introduction et/ou dans la liste des déclarations de brevets reçues par l'ISO (voir www.iso.org/brevets) ou dans la liste des déclarations de brevets reçues par l'IEC (voir patents.iec.ch).

Les appellations commerciales éventuellement mentionnées dans le présent document sont données pour information, par souci de commodité, à l'intention des utilisateurs et ne sauraient constituer un engagement.

Pour une explication de la nature volontaire des normes, la signification des termes et expressions spécifiques de l'ISO liés à l'évaluation de la conformité, ou pour toute information au sujet de l'adhésion de l'ISO aux principes de l'Organisation mondiale du commerce (OMC) concernant les obstacles techniques au commerce (OTC), voir: www.iso.org/iso/avant-propos.

Le présent document a été élaboré par le comité technique ISO/IEC/JTC 1, *Technologies de l'information*, sous-comité SC 2, *Jeux de caractères codés*.

Cette sixième édition annule et remplace la cinquième édition (ISO/IEC 14651:2019), qui a fait l'objet d'une révision technique.

Les principales modifications par rapport à l'édition précédente sont les suivantes:

— le présent document ajoute les données de classement des nouveaux caractères normalisés dans la sixième édition de l'ISO/IEC 10646 (2020) ;

— le contenu de 6.2.2 a été révisé pour le rendre plus complet ;

— les poids du caractère U+A9B5 (DIACRITIQUE VOYELLE JAVANAISE TOLONG) ont été modifiés, car ce dernier est considéré comme une variante du caractère U+A9B4 (DIACRITIQUE VOYELLE JAVANAISE TARUNG). Cela se devait d'être corrigé.

Il convient que l'utilisateur adresse tout retour d'information ou toute question concernant le présent document à l'organisme national de normalisation de son pays. Une liste exhaustive desdits organismes se trouve à l'adresse www.iso.org/members.html.

Introduction

Le présent document fournit une méthode universelle de mise en ordre des données textuelles. Elle fournit également une table-modèle commune qui, lorsqu'elle est adaptée, peut satisfaire aux exigences de tri d'une langue donnée, tout en triant de manière raisonnable les autres écritures.

La table-modèle commune est conçue de sorte qu'une adaptation s'avère nécessaire pour chaque environnement local. C'est pourquoi la conformité au présent document requiert que les modifications à cette table commune, appelées «deltas», soient déclarées de manière à documenter les différences dans les résultats.

Le présent document décrit une méthode pour classer l'information textuelle de manière indépendante du contexte.

L'ISO/IEC TR 30112 contient des dispositions pour le tri complémentaires à celles du présent document; on y trouve aussi des renseignements complémentaires sur les mots-clés définis dans le présent document et utilisés pour le tri.

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

1 Domaine d'application

Le présent document définit ce qui suit.

— Une méthode de référence pour la comparaison de deux chaînes de caractères ayant pour but de déterminer leur ordre de classement dans une liste triée. La méthode s'applique à des chaînes utilisant le répertoire complet de l'ISO/IEC 10646, des sous-répertoires tels que ceux des divers jeux normalisés ISO/IEC à 8 bits ou tout autre jeu de caractères, normalisé ou non, et permet de produire des résultats de tri valables (après adaptation) pour un ensemble de langues de chaque système d'écriture. Cette méthode de référence utilise des tables de tri dérivées soit de la table-modèle commune de classement définie dans le présent document, soit d'une de ses adaptations. La méthode procure un format de référence de la table-modèle commune. Ce format est décrit en notation BNF (forme de Backus-Naur, *Backus-Naur Form*). Son emploi est normatif *dans* le présent document.

— Une table-modèle commune de classement utilisée par la méthode de référence. Cette table décrit un ordre de base pour tous les caractères du standard Unicode 13.0[27] compris dans l'ISO/IEC 10646:2020. Tout cela permet de spécifier un ordre complètement déterministe. Cette table constitue le point de départ permettant de préciser un ordre de classement adapté aux règles de classement locales, sans qu'il soit nécessaire de connaître tous les systèmes d'écriture repris dans le jeu universel de caractères codés (JUC).

NOTE 1 Cette table-modèle commune de classement est destinée à être modifiée pour satisfaire aux besoins d'environnements locaux. L'avantage principal de cette pratique, sur le plan mondial, réside dans le fait que, pour d'autres systèmes d'écriture que celui de l'utilisateur, aucune modification n'est nécessaire et cet ordre demeurera aussi cohérent que possible et prévisible dans un contexte international.

NOTE 2 Le répertoire de caractères utilisé dans le présent document est équivalent à celui du standard Unicode, version 13.0[27].

— Un nom de référence représentant cette version particulière de la table-modèle commune, à utiliser comme point de départ à toute adaptation. Ce nom implique notamment que la table est liée à un stade de développement particulier du jeu universel de caractères codés (ISO/IEC 10646).

— Des exigences pour la déclaration de différences (delta) entre une table de tri et la table-modèle commune.

Le présent document *ne* spécifie *pas* ce qui suit.

— Une méthode particulière de comparaison; toute méthode équivalente conduisant aux mêmes résultats est acceptable.

— Un format précis pour décrire ou pour adapter les tables dans une mise en œuvre donnée.

— Des symboles précis à utiliser par les mises en œuvre, sauf pour ce qui est du nom de la table-modèle commune de classement.

— Une interface utilisateur particulière destinée à choisir les options.

— Un format interne particulier pour les clés intermédiaires utilisées dans les comparaisons ou pour la table de tri. L'utilisation de clés numériques n'est pas spécifiée non plus.

— Un ordre dépendant du contexte.

— Un prétraitement particulier des chaînes de caractères avant comparaison.

NOTE 1 Bien que ceci ne soit pas spécifié par le présent document, il s'avère souvent nécessaire de préparer les chaînes de caractères avant leur comparaison (cf. l'Annexe C).

NOTE 2 L'Annexe D décrit les problèmes qui ont donné lieu à la présente Norme internationale avec leurs solutions anticipées.

2 Références normatives

Les documents suivants sont cités dans le texte de sorte qu'ils constituent, pour tout ou partie de leur contenu, des exigences du présent document. Pour les références datées, seule l'édition citée s'applique. Pour les références non datées, la dernière édition du document de référence s'applique (y compris les éventuels amendements).

ISO/IEC 10646:2020, *Technologies de l'information — Jeu universel de caractères codés (JUC)*

3 Termes et définitions

Pour les besoins du présent document, les termes et définitions suivants s'appliquent.

L'ISO et l'IEC tiennent à jour des bases de données terminologiques destinées à être utilisées en normalisation, consultables aux adresses suivantes:

— ISO Online browsing platform: disponible à l'adresse https://www.iso.org/obp

— IEC Electropedia: disponible à l'adresse http://www.electropedia.org/

3.1
chaîne de caractères
suite de caractères considérée comme un objet simple

Note 1 à l'article: Note à l'article: Une chaîne de caractères à trier ne comprend normalement pas les caractères qui la délimitent, comme par exemple un caractère de commande de fin de ligne dans un fichier texte à trier.

3.2
symbole de tri
symbole (3.12) utilisé pour préciser les poids attribués à un *élément de tri* (3.4)

3.3
table de tri
table de poids
table reliant les *éléments de tri* (3.4) aux *éléments de poids* (3.14)

3.4
élément de tri
suite constituée d'un ou de plusieurs caractères considérés comme une seule entité aux fins du *tri* (3.7)

3.5
delta
liste des différences que présente une *table de tri* (3.3) donnée par rapport à une autre

Note 1 à l'article: Une table de tri donnée associée à un delta donné forme une nouvelle table de tri.

Note 2 à l'article: Sauf mention contraire, le terme «delta» désigne toujours les différences par rapport à la table-modèle commune définie dans le présent document.

3.6
niveau
niveau de tri
numéro d'une *sous-clé* (3.11) dans la série de sous-clés formant une clé

3.7
tri
procédé par lequel on détermine si, de deux chaînes, la première est plus petite, égale ou plus grande que la seconde

3.8
clé de tri
série de *sous-clés* (3.11) utilisée pour déterminer un ordre

3.9
prétraitement
procédé par lequel des *chaînes de caractères* (3.1) données sont transformées en d'autres chaînes avant le calcul de la *clé de tri* (3.8) de chaque chaîne

3.10
méthode de comparaison de référence
méthode de détermination de l'ordre relatif de deux *clés de tri* (3.8)

Note 1 à l'article: Voir l'Article 6.

3.11
sous-clé
suite de poids calculée pour une *chaîne de caractères* (3.1)

3.12
symbole
élément de tri (3.4)

3.13
poids
poids de tri
entier positif, utilisé dans les *sous-clés* (3.11), pour indiquer l'ordre relatif des *éléments de tri* (3.4)

3.14
élément de poids
liste d'un certain nombre de poids séquentiellement ordonnés par niveau

4 Symboles et conventions

Selon l'ISO/IEC 10646, les caractères se représentent à l'aide de U*X*, où *X* correspond à une série d'un à huit chiffres hexadécimaux (où toutes les lettres de la série de chiffres hexadécimaux sont en majuscules) et où *X* est le numéro du caractère dans l'ISO/IEC 10646. Cette convention est reprise dans le présent document.

Dans la table-modèle commune, des symboles arbitraires représentent des poids selon la notation BNF décrite en 6.3.1.

5 Conformité

Un processus est conforme au présent document s'il produit des résultats identiques à ceux qui résultent de l'application des spécifications décrites en 6.2 à 6.5.

Toute déclaration de conformité au présent document doit être accompagnée, directement ou par référence, d'une déclaration de ce qui suit.

— Le nombre de niveaux de tri que le processus peut utiliser; ce nombre doit être égal ou supérieur à trois.

— Si le paramètre de traitement forward,position est permis.

— Si le paramètre de traitement backward est permis et à quel niveau.

— Le *delta* d'adaptation décrit en 6.4 et le nombre de niveaux définis dans ce delta.

— Si un processus de prétraitement est utilisé, la méthode utilisée doit être déclarée.

Il incombe au producteur de montrer en quoi sa déclaration de delta est reliée à la syntaxe de la table décrite en 6.3, et comment la méthode de comparaison utilisée, si elle est différente de celle mentionnée à l'Article 6, peut être considérée comme produisant les mêmes résultats que ceux spécifiés par la méthode décrite à l'Article 6. L'usage d'un processus de prétraitement est optionnel et ses détails ne sont pas précisés dans le présent document.

Il est fortement recommandé que l'application présente à l'utilisateur les options et adaptations disponibles.

6 Comparaison de chaînes

6.1 Prétraitement des chaînes de caractères avant comparaison

Il peut s'avérer nécessaire de transformer les chaînes de caractères avant de leur appliquer la méthode de comparaison de référence. Bien que n'étant pas l'objet du présent document, le prétraitement peut être une partie importante du processus de tri. Voir l'Annexe C pour des exemples de prétraitement.

Les caractères de la chaîne d'entrée doivent être codés conformément à l'ISO/IEC 10646 (JUC) ou une correspondance à l'ISO/IEC 10646 doit être fournie si une autre forme de codage est utilisée.

Par conséquent, une partie importante de la phase préparatoire consiste à transformer les caractères d'un codage non-JUC à des caractères du JUC fournis en entrée à la méthode de comparaison. Cette tâche peut comprendre notamment le traitement correct de séquences d'échappement dans le codage original, la transformation de caractères sans attribution dans le JUC à des positions de code dans la zone privée et la transposition de caractères dans le cas de chaînes qui ne seraient pas stockées en ordre logique. Par exemple, dans le cas de codages arabes en ordre visuel, les caractères doivent être mis en ordre logique; dans le cas de certains codages à usage bibliographique, les accents combinatoires stockés avant leur caractère de base doivent être déplacés après le caractère de base. La suite résultante peut devoir être retransformée dans le codage original.

La table-modèle commune est conçue de sorte que les séquences combinatoires et les caractères simples (précomposés) correspondants aient exactement le même ordre. Pour éviter de violer par mégarde cet invariant (et au passage la conformité à Unicode), il convient que l'adaptation change le classement des séquences combinatoires quand le classement des caractères précomposés correspondants est changé. Par exemple, si Ä est déplacé après Z, il convient alors de changer aussi le classement de la séquence combinatoire <A>+ < tréma combinatoire >. Pour éviter de révéler des différences de codage invisibles à l'utilisateur, on recommande de normaliser les chaînes selon la forme FND de l'algorithme de normalisation Unicode – voir [29] dans la bibliographie.

Les séquences d'échappement et les caractères de commande sont très délicats à interpréter; il est fortement recommandé de les filtrer ou de les transformer.

NOTE Puisque la méthode de comparaison de référence est une description logique du procédé de comparaison de chaînes, rien n'empêche une mise en œuvre de cette méthode d'utiliser exclusivement un codage autre qu'un codage du JUC, pour autant que les résultats obtenus soient les mêmes que si la méthode de référence était utilisée.

6.2 Construction des clés et comparaison

6.2.1 Préliminaires

6.2.1.1 Hypothèses

La table de tri est une transformation des éléments de tri en éléments de poids. Pour chaque élément de poids, la table-modèle commune décrit quatre niveaux. L'adaptation peut augmenter ou réduire ce nombre de niveaux, mais pas à moins de trois.

NOTE Dans la table-modèle commune, les niveaux ont généralement les significations suivantes, bien que cet usage ne soit pas absolu:

Niveau 1: ce niveau correspond généralement au jeu de lettres de base pour une écriture alphabétique, au jeu de caractères courants pour une écriture idéographique ou syllabique.

Niveau 2: ce niveau correspond généralement aux diacritiques pouvant accompagner les caractères de base de chaque écriture. En certaines langues, les lettres accentuées sont considérées comme des lettres de base de l'alphabet et ne sont pas affectées par ce niveau, mais seulement par le premier niveau. En espagnol par exemple, le N TILDE est considéré comme une lettre de base de l'alphabet latin; par conséquent, une adaptation pour l'espagnol changera la définition de N TILDE de «le poids d'un N au premier niveau et le poids d'un TILDE au second niveau» à «le poids d'un N TILDE (entre N et O) au premier niveau et une indication de l'absence de diacritique au second niveau». Pour certains caractères, on prend également en compte des variantes de forme au second niveau, par exemple ß (la LETTRE MINUSCULE LATINE S DUR), qui est traitée comme un équivalent de ss au premier niveau mais s'en distingue traditionnellement au second niveau.

Niveau 3: ce niveau est généralement associé aux distinctions de casse (majuscules-minuscules) ou aux variantes de formes (comme la distinction entre hiragana et katakana).

Niveau 4: ce niveau est généralement consacré aux distinctions pondérales plus fines que celles des autres niveaux. Le dernier niveau (le quatrième dans la table-modèle commune) est souvent utilisé pour donner des poids additionnels à des caractères «spéciaux», c'est-à-dire des caractères qui ne sont pas normalement utilisés dans l'orthographe des mots d'une langue (ponctuation, vignettes, etc.), souvent appelés «ignorables» dans le contexte du tri informatique.

6.2.1.2 Propriétés de traitement

Une table de tri adaptée donnée possède des propriétés spécifiques de balayage et de classement. Ces propriétés peuvent avoir été changées par l'adaptation.

Une direction de balayage (vers l'avant ou vers l'arrière) pour chaque niveau est utilisée pour indiquer comment traiter la chaîne. La direction de balayage est une propriété globale de chaque niveau défini dans la table adaptée.

Si le dernier niveau est supérieur à trois, il existe une propriété optionnelle de ce niveau appelée l'option «position»: lorsqu'elle est active, une comparaison des positions numériques de chaque caractère «ignorable» dans les deux chaînes est effectuée, avant de comparer leurs poids. En d'autres mots, si deux chaînes sont équivalentes à tous les niveaux sauf le dernier, la chaîne contenant un caractère ignorable en position la plus basse est classée avant l'autre. Si les caractères ignorables ont les mêmes positions, alors leurs poids sont considérés jusqu'à ce qu'une différence soit trouvée. Le traitement correct de cette propriété optionnelle n'est pas nécessaire à la conformité au présent document.

NOTE La direction de balayage (vers l'avant ou vers l'arrière) n'est normalement pas reliée à la direction naturelle d'écriture. La direction de balayage s'applique à la suite logique de la chaîne de caractères codés.

Dans le cas d'écritures de droite à gauche comme l'arabe, l'ISO/IEC 10646 spécifie que les premiers caractères en ordre logique sont ceux apparaissant à droite en ordre de présentation. En écriture latine au contraire, les premiers caractères en ordre logique apparaissent à gauche en ordre de présentation.

Le balayage vers l'avant commence au début de la séquence en ordre logique, alors que le balayage vers l'arrière commence à la fin, sans égard à la direction de présentation. La direction de balayage à des fins de tri est une propriété globale de chaque niveau décrit dans la table.

Dans l'ISO/IEC 10646, l'écriture arabe est artificiellement séparée en deux pseudo-écritures: 1) l'écriture arabe logique, intrinsèque, codée indépendamment des formes contextuelles et 2) les formes de présentations arabes. Les deux permettent le codage complet de l'arabe, mais le codage intrinsèque est normalement privilégié pour sa meilleure capacité de traitement, alors que certaines applications de présentation préfèrent les formes de présentation. L'ISO/IEC 10646 ne spécifie pas l'ordre de stockage des formes de présentation; dans certaines réalisations, elles sont stockées en ordre inverse de celui utilisé pour le codage intrinsèque. Par conséquent, il est recommandé, lors de la phase de préparation, de s'assurer que les formes de présentation arabes et les autres caractères arabes soient fournis en ordre logique à la méthode de comparaison.

Une table de tri adaptée peut être séparée en sections pour faciliter l'adaptation. On donne alors à chaque section un nom, conformément aux dispositions de 6.3.1. Une des possibilités d'adaptation est de donner un certain ordre à chaque section et de changer l'ordre relatif d'une section par rapport à d'autres.

6.2.2 Méthode de référence de construction des clés

6.2.2.1 Généralités

Lorsque deux chaînes doivent être comparées pour déterminer leur ordre relatif, elles sont d'abord analysées en séquences d'éléments de tri, en tenant compte des déclarations «collating-element» à caractères multiples présentes dans la table de tri (*si* la syntaxe de 6.3.2 est utilisée). Dans la syntaxe utilisée pour exprimer la table-modèle commune, le nom d'un élément de tri associé à un seul caractère est formé de la lettre «U» suivie du numéro du caractère dans le JUC, en notation hexadécimale. Les noms et caractères associés aux éléments de tri multicaractère sont définis par les déclarations d'éléments de tri.

NOTE Les éléments de tri comportant plus de caractères sont préférés à ceux qui sont plus courts. Par exemple, si un élément de tri comportant plusieurs caractères est défini pour «abc» et qu'un autre est défini pour «ab» ou qu'un autre l'est pour «bc», alors, si «abc» se présente, l'élément de tri pour «abc» s'appliquera et non celui pour «ab» ou «bc».

Une suite de *m* sous-clés intermédiaires est alors formée de chaque chaîne, *m* étant le nombre de niveaux décrits dans une table de poids de tri adaptée.

Chaque clé de tri est une suite de sous-clés. Chaque sous-clé est une liste de poids numériques. Une sous-clé est construite en ajoutant successivement la liste des poids attribués à chaque élément de tri de la chaîne au niveau de la sous-clé en construction. Dans la table-modèle commune, le mot-clé «`IGNORE`» trouvé en place d'une suite de poids à un niveau indique que la suite de poids à ce niveau pour cet élément de tri est vide.

6.2.2.2 Annulation de certains éléments de tri

Dans la formation d'une clé de tri, tout élément de tri qui est ignoré au premier niveau ou aux deux premiers niveaux et qui succède à un élément de tri ignoré à tous les niveaux sauf le dernier ne conserve pas ses poids tels qu'ils sont donnés dans la table-modèle commune (ou son adaptation), mais chacun de ces poids doit être annulé (cela signifie que l'on doit remplacer chaque poids non nul par un poids nul: <IGNORE>).

6.2.2.3 Calcul de poids implicites

Si la table adaptée ne contient pas d'entrée pour un caractère de la chaîne d'entrée, les poids de ce caractère ne sont pas définis. Il faut dans ce cas calculer un poids primaire dit «implicite», constitué d'une paire de seizets – appelons-les «aaaa» et «bbbb» – et supposer l'existence de lignes d'adaptation de la forme suivante:

<UXXXX> "<R{aaaa$_{16}$}><T{bbbb$_{16}$}>";<BASE>;<MIN>;<SFFFF>

NOTE <SFFFF> (au dernier niveau) est le plus grand poids de premier niveau dans la table-modèle commune.

La règle de calcul d'un poids implicite n'est pas uniforme; des distinctions doivent être faites parmi les caractères qui n'ont pas d'entrée dans la table de tri:

a) Idéogrammes han unifiés:

Pour un caractère han à la position de code *pc*:

base_poids = 0xFB40 pour le RUO original

base_poids = 0xFB80 pour les caractères han des extensions A à G

aaaa = [base_poids + (pc >> 15)]$_{16}$

bbbb = [(pc & 0x7FFF) | 0x8000]$_{16}$

b) Caractères idéographiques et composants tangoutes:

Pour un caractère tangoute à la position de code *pc*:

base_poids = 0xFB00

aaaa = [base_poids]$_{16}$

bbbb = [(pc - 0x17000) | 0x8000]$_{16}$

c) Caractères idéographiques nüshu:

Pour un caractère nüshu à la position de code *pc*:

base_poids = 0xFB01

aaaa = [base_poids]$_{16}$

bbbb = [(pc - 0x1B170) | 0x8000]$_{16}$

d) Caractères idéographiques de la petite écriture khitane:

Pour un caractère de la petite écriture khitane à la position de code *pc*:

base_poids = 0xFB02

aaaa = [base_poids]$_{16}$

bbbb = [(pc - 0x18B00) | 0x8000]$_{16}$

e) Les autres points de code non mentionnés explicitement dans la table, c'est-à-dire les non-caractères, les demi-codets et tout point de code réservé pour une normalisation future (autrement dit, non encore attribué à un caractère du répertoire):

Pour un caractère à la position de code *pc*:

base_poids = 0xFBC0

aaaa = [base_poids + (pc >> 15)]$_{16}$

bbbb = [(pc & 0x7FFF) | 0x8000]$_{16}$

Les valeurs sont donc obtenues par le truchement de calculs d'opérations bit à bit, et dans chaque cas elles doivent être exprimées en base 16. Le poids de premier niveau se présente alors sous la forme suivante: "<Raaaa><Tbbbb>".

Les caractères décomposables sont exclus de ces traitements, car ils ont une entrée dans la table modèle commune (avec des poids de premier niveau sous la forme d'une paire couplée).

Dans le cas d'une suite mal formée d'octets, il y a deux options possibles: soit chaque octet de la séquence est ignoré, soit la séquence est traitée comme s'il s'agissait du caractère U+FFFD (CARACTÈRE DE REMPLACEMENT). Les mêmes options s'appliquent à d'éventuelles valeurs hors limite (pc > $10FFFF_{16}$).

NOTE 1 Les plages dans lesquelles sont codés les caractères han, tangoutes, nüshu et de la petite écriture khitane sont définies dans les commentaires ajoutés à la table-modèle commune (à la fin du fichier).

NOTE 2 Avec la méthode des poids implicites, les caractères sans entrée dans la table de tri sont ordonnés selon la valeur scalaire du point de code au sein de l'ensemble (han, tangoute, nüshu, petite écriture khitane, autre) auquel ils appartiennent, et ils sont triés correctement relativement aux autres caractères.

6.2.2.4 Formation des sous-clés

Il y a trois façons de former des sous-clés: vers l'avant (paramètre de traitement «forward»), vers l'arrière (paramètre de traitement «backward») et de façon positionnelle (paramètre de traitement «forward,position»). Les sous-clés formées de façon positionnelle ne peuvent apparaître qu'au dernier niveau et seulement si ce niveau est supérieur à trois. La conformité n'exige pas la formation de sous-clés de façon positionnelle; une réalisation incapable de formation positionnelle doit interpréter «forward,position» comme s'il s'agissait de «forward».

6.2.2.5 Formation des sous-clés aux trois premiers niveaux

En présence du paramètre de traitement «forward» ou «forward,position», on construit la sous-clé en balayant vers l'avant un à un les éléments de tri de la chaîne de caractères d'entrée pour leur attribuer un poids. On obtient les poids en recherchant les éléments de tri dans la table de poids de tri adaptée donnée et en extrayant la liste de poids pour le niveau considéré. Cette liste de poids s'ajoute à la fin de la sous-clé.

En présence du paramètre de traitement «backward» à un niveau donné, on construit la sous-clé vers l'avant, comme indiqué ci-devant, et on la renverse, poids par poids.

6.2.2.6 Formation de la sous-clé au quatrième niveau

Au dernier niveau, la sous-clé est construite comme cela est décrit dans l'alinéa précédent. Toutefois, une fois que la clé de tri est formée complètement (lorsque la fin de la chaîne est atteinte), deux options se présentent:

a) En présence du paramètre de traitement «forward»: on supprime *toutes* les occurrences du poids <SFFFF> qui se présentent dans la sous-clé;

b) En présence du paramètre de traitement «forward,position»: on doit supprimer de la sous-clé toute séquence de queue de la valeur <SFFFF> (la partie qui subsiste reste intacte).

6.2.3 Méthode de comparaison de référence pour le tri des chaînes de caractères

La méthode de comparaison de référence pour le classement de deux chaînes de caractères (*après* le prétraitement, qui ne fait pas partie de cette méthode de comparaison) consiste à comparer les clés de tri construites selon la méthode de référence décrite en 6.2.2 du présent document:

— En utilisant une table de poids de tri adaptée donnée, construire une clé de tri pour chacune des chaînes à comparer.

— Comparer ensuite les clés selon la définition de l'ordre des clés donnée en 6.2.4. Les clés peuvent être comparées jusqu'à un niveau donné ou jusqu'au dernier niveau de la table de poids de tri adaptée donnée.

NOTE La comparaison peut être effectuée *pendant* la construction des clés, en arrêtant cette construction dès que l'ordre des chaînes peut être déterminé. Cette technique est parfois appelée *évaluation paresseuse* et certains systèmes l'utilisent implicitement. Elle permet d'éviter la construction complète des clés quand une différence est trouvée tôt pendant la construction. Quand un ensemble important de chaînes doit être trié, l'on peut par exemple construire et stocker les clés – ou tout au moins un segment initial – avant de les comparer.

6.2.4 Méthode de comparaison de référence pour le tri des chaînes de caractères

Il convient de ne pas comparer les poids associés à des niveaux différents, ni par conséquent les sous-clés de différents niveaux. Il convient de ne pas comparer non plus les clés construites à partir de tables adaptées différentes.

NOTE 1 Ceci permet aux mises en œuvre d'attribuer les poids à chaque niveau indépendamment des autres niveaux et sans égard à d'autres tables adaptées.

m est le plus grand niveau d'une table adaptée donnée. Rappelons qu'une clé est une liste, de longueur m, de sous-clés; une sous-clé est une liste de poids; un poids est un entier positif. D'autres notations utilisées ci-dessous sont:

— L_z est la longueur de la sous-clé z, c'est-à-dire le nombre de poids dans cette sous-clé.

— $z_{\mathbf{pd}(a)}$, où $1 \le a \le L_z$, est le poids à la position a (un entier > 0) de la sous-clé z.

— $u_{\mathbf{sc}(b)}$, où $1 \le b \le m$, est la sous-clé de niveau b (un entier > 0) de la clé u.

Les ordres des poids, des sous-clés et des clés de tri (jusqu'à un certain niveau ou jusqu'au dernier niveau) sont des relations d'ordre total, définies pour une table de tri adaptée donnée comme suit:

a) Les poids sont des valeurs entières positives (dans la méthode de référence) et sont comparés comme tels aux fins du classement.

b) Une sous-clé v est *plus petite* qu'une sous-clé w (on notera $v < w$) **si et seulement s'il** existe un entier i, où $1 \le i \le L_v+1$ et $i \le L_w$, tel que

   — $i = 1$ et $v_{\mathbf{pd}(i)} < w_{\mathbf{pd}(i)}$, ou

   — pour tous les entiers j, $1 \le j < i$, l'égalité $v_{\mathbf{pd}(j)} = w_{\mathbf{pd}(j)}$ est maintenue, et soit

      — $i \le L_v$ et $v_{\mathbf{pd}(i)} < w_{\mathbf{pd}(i)}$, soit

      — $i = L_v+1$ et $0 < w_{\mathbf{pd}(i)}$.

   Une sous-clé v est *plus grande* qu'une sous-clé w (on notera $v > w$) **si et seulement si** w est plus petite que v. Une sous-clé v est *égale* à une sous-clé w (on notera $v = w$) **si et seulement si** v n'est pas plus petite que w et w n'est pas plus petite que v.

c) Une clé de tri x est *plus petite* qu'une clé de tri y au niveau s (on notera $x <_s y$) **si et seulement s'il** existe un entier i, où $1 \le i \le s$ et $i \le m$, tel que

   — $i = 1$ et $x_{\mathbf{sc}(i)} < y_{\mathbf{sc}(i)}$, ou

   — pour tous les entiers j, $1 \le j < i$, l'égalité $x_{\mathbf{sc}(j)} = y_{\mathbf{sc}(j)}$ est maintenue, et $x_{\mathbf{sc}(i)} < y_{\mathbf{sc}(i)}$.

   Une clé de tri x est *plus grande* qu'une clé de tri y au niveau s (on notera $x >_s y$) **si et seulement si** y est plus petite que x au niveau s. Une clé de tri x est *égale* à une clé de tri y au niveau s (on notera $x =_s y$) **si et seulement si** x n'est pas plus petite que y au niveau s et y n'est pas plus petite que x au niveau s.

d) Pour les clés de tri, $<$, $>$ et $=$ sont définis comme $<_m$, $>_m$ et $=_m$ respectivement.