INTERNATIONAL STANDARD

ISO/IEC 14651

First edition
2001-02-15

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

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

Reference number
ISO/IEC 14651:2001(E)

PDF disclaimer

This PDF file may contain embedded typefaces. In accordance with Adobe's licensing policy, this file may be printed or viewed but shall not be edited unless the typefaces which are embedded are licensed to and installed on the computer performing the editing. In downloading this file, parties accept therein the responsibility of not infringing Adobe's licensing policy. The ISO Central Secretariat accepts no liability in this area.

Adobe is a trademark of Adobe Systems Incorporated.

Details of the software products used to create this PDF file can be found in the General Info relative to the file; the PDF-creation parameters were optimized for printing. Every care has been taken to ensure that the file is suitable for use by ISO member bodies. In the unlikely event that a problem relating to it is found, please inform the Central Secretariat at the address given below.

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.ch
Web www.iso.ch

Printed in Switzerland

Contents

Foreword iv

Introduction v

1. Scope 1

2. Conformance 2

3. Normative references 2

4. Terms and definitions 3

5. Symbols and abbreviations 4

6. String comparison 4

6.1. Preparation of character strings prior to comparison 4
6.2. Key building and comparison 5
6.2.1. Preliminary considerations 5
6.2.2. Reference ordering key formation 6
6.2.3. Reference comparison method for ordering character strings 7
6.3. Common Template Table: formation and interpretation 8
6.3.1. BNF syntax rules for the Common Template Table in Annex A 9
6.3.2. Well-formedness conditions 12
6.3.3. Interpretation of tailored tables 13
6.3.4. Evaluation of weight tables 14
6.3.5. Conditions for considering specific table equivalences 15
6.3.6. Conditions for results to be considered equivalent 15
6.4. Declaration of a delta 16
6.5. Name of the Common Template Table and name declaration 17

Annex A (normative) Common Template Table 18

Annex B (informative) Example tailoring deltas 19

B.1. Example 1 – Minimal tailoring 19
B.2. Example 2 – Reversing the order of lowercase and uppercase letters 19
B.3. Example 3 – Canadian delta and benchmark 19
B.4. Example 4 – Danish delta and benchmark 22

Annex C (informative) Preparation 25

C.1. General considerations 25
C.2. Thai string ordering – a case involving preparation required to get the proper ordering 25
C.2.1. Thai Ordering Principle 25
C.2.2 Algorithmic Aspect 27
C.3. Handling of numeral substrings in collation 29
C.3.1. Handling of 'ordinary' numerals for natural numbers 29
C.3.2. Handling of positional numerals in other scripts 32
C.3.3. Handling of other non-pure positional system numerals or non-positional system numerals (e.g. Roman numerals) 32
C.3.4. Handling of numerals for whole numbers 32
C.3.5. Handling of positive positional numerals with fractional parts 34
C.3.6. Handling of positive positional numerals with fraction parts and exponent parts 35
C.3.7. Handling of date and time of day indications 35
C.3.8. Making numbers less significant than letters 37
C.3.9. Maintaining determinacy 37

Annex D (informative) Tutorial on solutions brought by this standard to problems of lexical ordering 38

D.1. Problems 38
D.2. Solution 39
D.3. Tailoring 41

Bibliography 42

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.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 3.

In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1. 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 International Standard may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

International Standard ISO/IEC 14651 was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 22, *Programming languages, their environments and system software interfaces*.

Annex A forms a normative part of this International Standard. Annexes B, C and D are for information only.

Introduction

This International Standard provides a method, applicable around the world, for ordering text data, and provides a Common Template Table which, when tailored, can meet a given language's ordering requirements while retaining reasonable ordering for other scripts.

The Common Template Table requires some tailoring in different local environments. Conformance to this International Standard requires that all deviations from the Template, called "deltas", be declared to document resultant discrepancies.

This International Standard describes a method to order text data independently of context.

The Draft Technical Report ISO/IEC DTR 14652 (under development) has specifications for ordering that informatively complements the specifications in this International Standard, and where additional information may be sought on ordering keywords defined in this International Standard.

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

1. Scope

This International Standard defines:

- A reference comparison method. This method is applicable to two character strings to determine their relative order. The method can be applied to strings containing characters from the full repertoire of ISO/IEC 10646-1. This method is also applicable to subsets of that repertoire, such as those of the different ISO/IEC 8-bit standard character sets, or any other character set, standardised or not, to produce ordering results valid (after tailoring) for a given set of languages for each script. This method uses collation tables derived either from the Common Template Table defined in this International Standard or from one of its tailorings.
- A reference format. The format is described using the Backus-Naur Form (BNF). This format is used to describe the Common Template Table. The format is used normatively *within* this International Standard.
- A Common Template Table. A given tailoring of the Common Template Table is used by the reference comparison method. The Common Template Table describes an order for all characters encoded in the first edition of ISO/IEC 10646-1 up to Amendment 7. It allows for a specification of a fully deterministic ordering. This table enables the specification of a string ordering adapted to local ordering rules, without requiring an implementer to have knowledge of all the different scripts already encoded in the UCS.

NOTE 1: This Common Template Table is to be modified to suit the needs of a local environment. The main benefit, worldwide, is that for other scripts, often no modification may be required and that the order will remain as consistent as possible and predictable from an international point of view.

NOTE 2: The character repertoire used in this International Standard is equivalent to that of the Unicode Standard version 2.1.

- A reference name. The reference name refers to this particular version of the Common Template Table, for use as a reference when tailoring. In particular, this name implies that the table is linked to a particular stage of development of the ISO/IEC 10646 Universal multiple-octet coded character set.
- Requirements for a declaration of the differences (delta) between the collation table and the Common Template Table.

This International Standard does *not* mandate:

- A specific comparison method; any equivalent method giving the same results is acceptable.
- A specific format for describing or tailoring tables in a given implementation.
- Specific symbols to be used by implementations except for the name of the Common Template Table.
- Any specific user interface for choosing options.
- Any specific internal format for intermediate keys used when comparing, nor for the table used. The use of numeric keys is not mandated either.

- A context-dependent ordering.
- Any particular preparation of character strings prior to comparison.

NOTE 1: It is normally necessary to do preparation of character strings prior to comparison even if it is not prescribed by this International standard (see informative Annex C).

NOTE 2: Although no user interface is required to choose options or to specify tailoring of the Common Template Table, conformance requires always declaring the applicable delta, a declaration of differences with this table. It is recommended that processes present available tailoring options to users.

2. Conformance

A process is conformant to this International Standard if it meets the requirements prescribed in subclauses 6.2 to 6.5.

A declaration of conformity to this International Standard shall be accompanied by a statement, either directly or by reference, of the following:

- The number of levels that the process supports; this number shall be at least three.
- Whether the process supports the `forward,position` processing parameter.
- Whether the process supports the `backward` processing parameter and at which level.
- The tailoring *delta* described in clause 6.4 and how many levels are defined in the delta.

It is the responsibility of implementers to show how their delta declaration is related to the table syntax described in 6.3, and how the comparison method they use, if different from the one mentioned in clause 6, can be considered as giving the same results as those prescribed by the method specified in clause 6.

3. Normative references

The following normative documents contain provisions which, through reference in this text, constitute provisions of this International Standard. For dated references, subsequent amendments to, or revisions of, any of these publications do not apply. However, parties to agreements based on this International Standard are encouraged to investigate the possibility of applying the most recent editions of the normative documents indicated below. For undated references, the latest edition of the normative document referred to applies. Members of ISO and IEC maintain registers of currently valid International Standards.

- ISO/IEC 10646-1:1993 *Information technology – Universal Multiple-Octet Coded Character Set (UCS) – Part 1: Architecture and Basic Multilingual Plane*.
- ISO/IEC 10646-1:1993/Amd.1:1996 *Information technology – Universal Multiple-Octet Coded Character Set (UCS) – Part 1: Architecture and Basic Multilingual Plane Amendment 1: Transformation Format for 16 planes of group 00 (UTF-16)*.
- ISO/IEC 10646-1:1993/Amd.2:1996 *Information technology – Universal Multiple-Octet Coded Character Set (UCS) – Part 1: Architecture and Basic Multilingual Plane Amendment 2: UCS Transformation Format 8 (UTF-8)*.
- ISO/IEC 10646-1:1993/Amd.4:1996 *Information technology – Universal Multiple-Octet Coded Character Set (UCS) – Part 1: Architecture and Basic Multilingual Plane Amendment 4*.

- ISO/IEC 10646-1:1993/Amd.5:1998 *Information technology – Universal Multiple-Octet Coded Character Set (UCS) – Part 1: Architecture and Basic Multilingual Plane Amendment 5: Hangul syllables*.
- ISO/IEC 10646-1:1993/Amd.6:1997 *Information technology – Universal Multiple-Octet Coded Character Set (UCS) – Part 1: Architecture and Basic Multilingual Plane Amendment 6: Tibetan*.
- ISO/IEC 10646-1:1993/Amd.7:1997 *Information technology – Universal Multiple-Octet Coded Character Set (UCS) – Part 1: Architecture and Basic Multilingual Plane Amendment 7: 33 additional characters*.
- ISO/IEC 10646-1:1993/Amd.9:1997 *Information technology – Universal Multiple-Octet Coded Character Set (UCS) – Part 1: Architecture and Basic Multilingual Plane Amendment 9: Identifiers for characters*.
- ISO/IEC 10646-1:1993/Amd.18:1999 *Information technology – Universal Multiple-Octet Coded Character Set (UCS) – Part 1: Architecture and Basic Multilingual Plane Amendment 18: Symbols and other characters*.

NOTE: Only the EURO SIGN and the OBJECT REPLACEMENT CHARACTER from Amendment 18 are accounted for in Annex A of this International Standard at this time.