TECHNICAL REPORT

RAPPORT TECHNIQUE

TECHNISCHER BERICHT

CEN/TR 14381

April 2003

ICS 35.040

English version

Information technology – Character repertoire and coding transformations – European fallback rules

This Technical Report was approved by CEN on 2 June 2002. It has been drawn up by the Technical Committee CEN/TC 304.

CEN members are the national standards bodies of Austria, Belgium, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Luxembourg, Malta, Netherlands, Norway, Portugal, Slovakia, Spain, Sweden, Switzerland and United Kingdom.

EUROPEAN COMMITTEE FOR STANDARDIZATION
COMITÉ EUROPÉEN DE NORMALISATION
EUROPÄISCHES KOMITEE FÜR NORMUNG

Management Centre: rue de Stassart, 36 B-1050 Brussels

Ref. No. CEN/TR 14381:2003 E

Contents

Foreword

This document (CEN/TR 14381:2003) has been prepared by Technical Committee CEN/TC 304 "Information and communications technologies - European localization requirements", the secretariat of which is held by IST.

The text of this technical report was written with the intent of it being published as a European pre-Norm (ENV). In light of the various formal and informal comments received on the document (some of which were only received after the closing of the ballot) the TC has resolved to turn this document into a CEN report as a recorded example of an attempt to formulate European wide fallback rules. It is evident that any fallback scheme in order for it to become acceptable by the users and the industry will need to be very carefully laid out and explained.

Resolutions no 7 of the 16th Meeting and no 4 of the 18th Meeting of CEN/TC 304 refer to this technical report:

> Res. 7/16th. TC304 acknowledges that the Fallback project team has completed its contracted work. Although the proposed draft has received the sufficient support to be forwarded to CEN/BT for final adoption as an ENV, the nature of the comments received is such that it is decided to publish it as a CEN Report with editorial comments added by the secretary and reviewed by TC members and observers before final publication. Unanimous.
>
> Res. 4/18th. TC304 accepts the Fallback document in N978 to be presented to CEN BT for adoption as a CR with the following text on Greek letters added in the foreword: “The method of performing fallback from Greek letters into Latin letters is especially seen as posing problems to Greek users and its use is not advised". Unanimous.

This technical report is intended to facilitate cross border communications and data exchange and to ensure that European cultural requirements are safeguarded in the increasingly interconnected world of today. It provides rules for fallback for multilingual European texts into the invariant set of ISO/IEC 646. These rules come into effect if data from different languages must be represented by equipment and systems that do not support the presentation of all the characters in the different language repertoires.

This technical report does not intend to influence, let alone substitute itself for, national standards or customs in this field. Nevertheless, national standards have the opportunity to adapt this Technical Report by declaring a formalized set of deviation rules (»delta«) if they so wish.

This document does not cancel or replace any other technical report or standard.

There is no known identical national technical report or standard in Europe.

According to the CEN/CENELEC Common Rules the following countries are bound to announce the existence of this Technical Report: Austria, Belgium, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Luxembourg, Malta, Netherlands, Norway, Portugal, Slovakia, Spain, Sweden, Switzerland and United Kingdom.

0 Introduction

0.1 Rationale for the provision of fallback rules

Users who are trying to write text in a language which is not their mother tongue (native language) often wish to write that text using a character repertoire which does not contain all the letters needed for that language, especially those with diacritic marks. A method of character substitution would be useful for such users.

In spite of the computers being able to process larger repertoires of graphic characters than ever before, there are cases where it is not possible to render all the characters of a processing repertoire on an output device. In these cases, not all the characters in a processing repertoire are available in an output repertoire. In order to cater for these situations, a widely applicable standard method of character substitution (fallback) is required which will allow an approximate rendition to be made of the unsupported characters of the processing repertoire for output and rendition.

Examples of key applications are:

a) a multilingual information service offered across Europe where personal or business documents come from different countries and are presented in a standardised rendition using MES 2 characters and which cannot be properly represented by the information service; and

b) search engines in the World Wide Web which make use of "fuzzy" search techniques based on the use of search terms which have diacritical marks removed and make use of common substitutions for less frequently used letters of the Latin alphabet. Examples of the latter are - eth (ð), thorn (Þ), æ, œ and the German sharp s (ß)

The provision of single fallback rules with a collection of fallback representations for MES 2 will enable the services to be improved and the applications easier for use by the human end user. A standard set of substitutions would be useful for such applications in order to avoid confusion. The same applies for the other two scripts represented by MES-2, Cyrillic and Greek.

The justification for preparing a technical report for these purposes is that the concept of representing the characters without diacritical marks is not useful for scripts originating outside Europe. Furthermore, Standardisation bodies of Europe that may wish to specify national schemes for fallback may modify the scheme given in this technical report for a limited set of characters and promulgate national standards for fallbacks. Greek and Cyrillic fallback representation specified in this technical report should be used with caution since transliteration into various Latin script languages depends on the target language. Local standards or local best practice should be referenced where they exist.

This European fallback specification can be used as a default in all relevant situations. It can be used as the basis for national standards with local preferences being used for specific substitutions defined by particular nation. It is expected that national standards for fallback will be registered in the international cultural registry as part of national locales. Well known local solutions will also be documented in addition to the default values.

0.2 Basic concepts

This standard specifies how a source stream of coded characters from a processing repertoire is represented in a target stream of an output repertoire. The worst case that is covered by the substitutions defined in this technical report is where the processing repertoire is MES-2 and the output repertoire is the invariant repertoire of ISO/IEC 646. The coding of the processing repertoire and the coding of the output repertoire are outside the scope of this TR.

Characters in the source stream that occur in the output repertoire are transferred directly to the target stream without substitution. Characters in the source stream that do not occur in the output repertoire are subject to substitution.

There are two types of substitution. In the first type, the target characters are represented in a way that disables the reverse transformation of the target stream to the source stream because of **loss of information**. A very common example of this type of presentation is when the Latin small letter e acute (é) is presented by Latin small letter e (e). This type of substitution when a letter with diacritical mark is represented with the same character but without diacritical mark is known as accent dropping. The second type of presentation introduces special symbols that preserve the information about the original graphical symbol enabling transformation of the character stream to the original encoding. An example of this is the use of the SGML symbols (e.g. *&eacute* in the case above). This type of substitution is outside the scope of this TR.

The substitution with loss of information can have more forms, but two main classes are always recognised as basic:

-one-to-many when one graphical character of the source stream is substituted with more than one graphical character from the output repertoire in the target stream. An example of this type of presentation is Latin capital letter Æ presented as AE. This class is recommended for general use.

-one-to-one when one graphical character of the source stream is substituted with one graphical character of the output repertoire in the target stream. This type of presentation is required in applications were the number of characters in the data entries or fields (e.g. in data bases or application forms) is fixed. The accent dropping is a type of one to one substitution. It is anticipated that this class will have minority application and should be discouraged, only to be used when there are strong technical reasons for doing so.

0.3 Requirements

It is desirable that a standard fallback specification has a very large field of application so that it may be used across a wide range of platforms. To achieve this, a fallback specification is needed for a processing repertoire that is a superset of a large proportion of existing processing repertoires. Also, a fallback repertoire is needed which is a subset of a large proportion of existing output repertoires.

The characters contained in the collection Multilingual European Subset No.2 (MES-2) specified in CEN CWA 13873-2000, have a wide usage across Europe. In many cases, MES-2 will become the processing repertoire of choice. MES-2 is a large repertoire for which there will be a need for a fallback specification.

0.4 Satisfying the requirements

MES-2 has been designed to be a superset of a wide range of processing repertoires for commercial and administrative applications in office environments across the EEA.

It is a valid assumption that the minimum output repertoire that is implemented in computer systems is the invariant repertoire of ISO/IEC 646. Therefore a substitution with one or more characters from the invariant repertoire of ISO/IEC 646 will always be able to be rendered on an output device. The invariant repertoire of ISO/IEC 646 does not have letters with diacritic marks nor letters from national alphabets used in Europe.

This standard satisfies the requirements by providing a fallback specification that can represent each character of MES-2 (and six additional characters). The additional six characters are 048C, 048D, 048E, 048F, 04EC and 04ED.

1. Scope and field of application

1.1 Scope

This technical report specifies the representation of the characters of the collection Multilingual European Subset No.2, MES-2, and six additional characters with one or more characters of the invariant repertoire of ISO/IEC 646 (83 graphic characters). Where a character is not available in the invariant set of ISO/IEC 646, a fallback representation for rendering is specified.

1.2 Field of application

The fallback rules given here are only intended for data in more than one European language, i.e. for use in pan-European applications. They are not meant to influence, let alone replace existing national standards or practices.

2. Normative references

2.1 ISO/IEC 10646-1:2000, *Information technology-Universal Multiple-Octet Coded Character Set (UCS)-Part 1: Architecture and Basic Multilingual Plane.*

2.2 ISO/IEC 646:1991, *Information technology - ISO 7-bit coded character set for information interchange.*

3. Definitions and abbreviations

3.1 Basic definitions

For the purposes of this technical report the basic definitions of ISO/IEC 10646-1 section 4 apply. The following are reproduced here for ease of reference:

3.1.1 *character*: A member of a set of elements used for organisation, control, or representation of data.

3.1.2 *coded character*: A character together with its coded representation.

3.1.3 *coded character set*: A set of unambiguous rules that establishes a character set and the relationship between the characters of the set and their coded representation.

3.1.4 *repertoire*: A specified set of characters that are represented in a coded character set

3.1.5 *presentation*: To present: the process of writing, printing, or displaying of a graphic symbol.

3.1.6 *graphic symbol*: The visual representation of a graphic character or of a composite sequence.

3.1.7 *graphic character*: A character, other than a control function, that has a visual representation, normally hand-written, printed, or displayed.

3.2 Other definitions

Also, the following definitions apply:

3.2.1 *diacritical mark*: Any of a number of recurring graphical structures placed over, under, next to, or through a basic letter which does not significantly modify the shape of the basic letter itself and which in combination with that basic letter is a graphic character of the identified repertoire of MES-2.

3.2.2 *letter with diacritical mark*: A letter which is constructed as the combination of a basic letter and a diacritical mark.

3.2.3 *basic letter*: A letter that is one of the letters of the repertoire of the IRV of ISO 646.

3.2.4 *character stream*: A series of coded characters in sequence, sometimes referenced as a stream.