CEN REPORT

RAPPORT CEN

CEN BERICHT

CR 13907

May 2000

ICS

English version

Information Technology - Character Repertoire and Coding Transformations - General model for graphic character transformations

This CEN Report was approved by CEN on 10 April 2000. 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, Iceland, Ireland, Italy, Luxembourg, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland and United Kingdom.

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

Central Secretariat: rue de Stassart, 36 B-1050 Brussels

Ref. No. CR 13907:2000 E

FOREWORD

This report was produced by a CEN/TC304 Project Team, set up in June 1998, as one of several to carry out the funded work program of TC304 (documented in CEN/TC304 N666R2). This is the third draft, taking into account comments received at the TC meetings in Brussels in November 1998 and in Tübingen in April 1999.

TABLE OF CONTENT

Character Repertoire and Coding Transformations — General model for graphic character transformations

1 Scope

1.1 General principles

This Technical Report describes a general model of the conceptual stages involved in the interchange of data composed of graphic characters between two end users. It identifies those aspects of this communication process that are amenable to further standardization and it provides terminology that permits such standards to specify their roles within this model. It is not intended as a guide to the implementation of such standards as in many cases the conceptual stages do not correspond to the practical stages involved in an efficient implementation.

The general model addresses both situations in which the intention is the interchange of data without alteration and situations in which transformation of the data during interchange is either required or acceptable. Examples of the latter situation are required transliteration and acceptable fallback.

The general model covers both transformations that affect the character content of the data and those that affect its coded representation. It addresses in particular the issues that arise when some system involved in the communication process is unable to handle all the characters that the end users wish to convey. The general model is not concerned with the meaning of the character data that is being communicated, such as its language, or with its rendition attributes such as font, size and weight. The general model is also applicable when the interchange takes place within a single system with the primary aim of data transformation, such as transliteration or language translation.

1.2 Character environments

A character is an abstract concept which is represented in different ways in different environments. To identify the character corresponding to some particular representation, it is necessary to know the environment concerned. The glyph, or symbol, ‘A’ is no more the character with name LATIN CAPITAL LETTER A than is the hexadecimal code value ‘41’.
The former represents this character in the Latin script and the latter does so in the registered character set ISO-IR 6 (ASCII), but they equally represent GREEK CAPITAL LETTER ALPHA in the Greek script and ISO-IR 150 respectively, and CYRILLIC CAPITAL LETTER A in the Cyrillic script and ISO-IR 146 respectively.

The general model for graphic character transformations concerns the processes that are involved in the interchange of graphic character data between two users by means of a transport process that provides a transparent transfer of binary data. The model may be applied to transformation within a single system by treating the binary transport process as an internal interface. The model identifies the following hierarchy of environments as being involved in the communication process:

a) User environment;

b) Application environment;

c) Interchange environment.

In a user environment, characters are normally presented as glyphs. In both application and interchange environments, characters are represented by bit combinations according to some encoding scheme. In an interchange of character data, in general both the user environment and the application environment of the receiving system will differ from the corresponding environments of the sending system, but both systems will generally have the same interchange environment. The underlying transport process is then used to transfer between the two systems those bit combinations that represent characters in their common interchange environment. If the interchange environments differ, transparent transfer of binary data will not result in transparent transfer of character data between the

two environments. Other elements within the end systems concerned may, however, be able to compensate for the distortion of the character data so produced.

1.3 Code characteristics

The difference between application and interchange environments lies in the encoding schemes that are used. In an application environment it is potentially possible to use every available bit combination to represent a graphic character, so that the SPACE character and 255 other graphic characters can all be encoded in an 8-bit code. As an example, this potential is in fact realized in proprietary PC code pages. In such an environment, formatting and other control information is recorded separately from the graphic character data so that there is no need to reserve any code positions for control data. In an interchange environment, however, it is normally necessary to encode graphic characters and control characters together in a single binary stream. This requirement leads to the use in interchange environments of coded character sets in which certain code positions are reserved for control characters. The 8-bit codes used in such an environment generally follow the character code structure specified in ISO/IEC 2022, which reserves the hexadecimal code positions 00–1F and 80–9F for control characters with 20 being allocated to the SPACE character and 7F to the DELETE character. This leaves only 190 code positions available for graphic characters other than SPACE. The various parts of ISO/IEC 8859 specify coded graphic character sets with this structure, all of which have a common assignment of the ‘left hand’ part, *i.e.* of the code positions 20–7E, in accordance with ISO-IR 6 (ASCII).

The application environment, however, normally has a requirement for fixed-length codes, *i.e.* coded character sets in which every character is represented by the same number of bits.
Such codes simplify random access of stored data since the location of each coded character within a sequentially stored sequence is independent of the previous characters in the sequence. The interchange environment has no such requirement, so permitting characters with diacritical marks, for example, to be coded by the addition of further bits to the bit pattern that represents the base character. In this way the coded character set specified in ISO/IEC 6937 encodes 333 graphic characters including SPACE, yet it uses the 8-bit code values 20–7E for the same characters as does ISO/IEC 8859.

1.4 Modelling the environments

The differences in the natures of the user, application and interchange environments lead to differences in the character repertoires that they are capable of representing. This in turn leads to difficulties when character data is passed sequentially from one environment to another in the course of its transmission from one end user to another. The above descriptions of the different environments are, however, purely illustrative. The general model described in this Technical Report recognises the existence of these environments and the differences in their repertoires but the detailed features of the environments that lead to these repertoire differences are outside the scope of the report. In particular, the character encoding used in the application environment is outside the scope of the report; it is only the repertoire of this environment that enters the general model. The character encoding used in the interchange environment, however, is within the scope of the model since it is this encoding that provides the transformation from characters in the interchange environment to the binary data transferred by the transport process.

2 Definitions

For the purposes of this Technical Report, the following definitions apply. Unless otherwise specified, where a definition is followed by reference to an International Standard, it has been taken verbatim from that standard.

application environment: A system environment in which characters are represented by bit combinations for the purposes of an application process.

application process: An element within a real system which performs the information processing for a particular application.