SLOVENSKI STANDARD
SIST-TP CR 13907:2003

01-oktober-2003

Informacijska tehnologija ï Nabor znakov in kodirne pretvorbe ï Sploġni model za pretvarjanje grafiļ nih znakov

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

Ta slovenski standard je istoveten z: CR 13907:2000

ICS:

35.040	Nabori znakov in kodiranje informacij	Character sets and information coding

SIST-TP CR 13907:2003 en

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

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.

NOTE: This differs from the definition in ISO/IEC 7498-1:1994, in which "real system" is strengthened to "real open system", since the real system here need not comply with the requirements of OSI standards in its communication with other real systems.

bit combination: An ordered set of bits used for the representation of characters. [ISO/IEC 2022:1994]

byte: A bit string that is operated upon as a unit. (Note: Each bit has the value either ZERO or ONE.) [ISO/IEC 2022:1994]

CC-data-element (Coded-Character-Data-Element): An element of interchanged information that is specified to consist of a sequence of coded representations of characters, in accordance with one or more identified standards for coded character sets. [ISO/IEC 2022:1994, ISO/IEC 10646-1:1993]

character: A member of a set of elements used for the organization, control, or representation of data. [ISO/IEC 2022:1994, ISO/IEC 10646-1:1993]

character string: A sequence of characters selected from a specified repertoire.

character transformation: A process which maps character strings of some source repertoire into character strings of a target repertoire. A character transformation yields a unique string of characters in the target repertoire from every string of characters in the source repertoire. It need not, however, act on the source string on a character-by-character basis and it need not preserve the number of characters in the string. It may, but need not, be reversible.

NOTE: Both transcription and transliteration are character transformations in this sense, as would be any deterministic scheme of language translation.

coded character: A character together with its coded representation. [ISO/IEC 10646-1:1993]

coded character set; code: A set of unambiguous rules that establishes a character set and the one-to-one relationship between the characters of the set and their bit combinations. [ISO/IEC 2022:1994]

combining character: A member of an identified subset of a coded character set, intended for combination with the preceding or following graphic character, or with a sequence of combining characters preceded or followed by a non-combining character. [ISO/IEC 2022:1994]

composite sequence: A sequence of graphic characters consisting of a non-combining character followed by one or more combining characters. [ISO/IEC 10646-1:1993]

device: A component of information processing equipment which can transmit and/or receive coded information within CC-data-elements. (It may be an input/output device in the conventional sense, or a process such as an application program or gateway function.) [ISO/IEC 10646-1:1993]

environment: A correspondence between the characters of a specified repertoire and a set of objects used for the presentation or representation of those characters; examples of possible objects are glyphs, bit combinations and Braille patterns.

glyph: A recognisable abstract graphic symbol which is independent of any specific design. [ISO/IEC 9541-1:1991]

graphic character: A character, other than a control function, that has a visual representation normally handwritten, printed, or displayed. [ISO/IEC 10646-1:1993]

graphic symbol: The visual representation of a graphic character or of a composite sequence. [ISO/IEC 10646-1:1993]

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

interchange: The transfer of character coded data from one user to another, using telecommunication means or interchangeable media. [ISO/IEC 10646-1:1993]

presentation; to present: The process of writing, printing or displaying a graphic symbol. [ISO/IEC 10646-1:1993]

real system: A set of one or more computers, the associated software, peripherals, terminals, human operators, physical processes, information transfer means, etc. that forms an autonomous whole capable of performing information processing and/or information transfer. [ISO/IEC 7498-1:1994]

repertoire: A specified set of characters that are each represented by one or more bit combinations of a coded character set. [ISO/IEC 2022:1994]

NOTE: The characters contained in a repertoire need not all be represented in the same coded character set. The code extension techniques of ISO/IEC 2022, in particular, permit the construction of a CC-data-element which makes use of coded representations selected from any number of coded character sets and which therefore can represent the characters of any repertoire.

script: A set of graphic characters used for the written form of one or more languages. [ISO/IEC 10646-1:1993]

system environment: An environment determined by the capabilities of the computers, software, human operators etc. that form part of a real system. The system environments of a particular real system describe the maximum capabilities of that system to handle characters for one or more of input, output and processing, as appropriate. Where the same environment is concerned with more than one of these aspects of a real system, it describes the common capabilities of all aspects with which it is concerned.

transcription: The process whereby the pronounciation of a given language is noted by the system of signs of a conversion language. A transcription system is of necessity based on the orthographical conventions of the conversion language. Transcription is not strictly reversible. [ISO 3602:1989]

transliteration: The process which consists of representing the characters of an alphabetical or syllable writing system by the the characters of a conversion alphabet. In principle, this conversion should be made character by character. [ISO 3602:1989]

NOTE: Transliteration is a reversible process.

user: A person or other entity that invokes the service provided by a device. (This entity may be a process such as an application program if the "device" is a code converter or a gateway function, for example.) [ISO/IEC 10646:1993]

user environment: A system environment in which characters are represented by, or presented as, objects capable of identification by a particular user; examples for such objects are glyphs when the user is a person or an optical character reader, and bit combinations when the user is an application process.

3 Abbreviations

For the purposes of this Technical Report, the following abbreviations apply.

ASCII	American Standard Code for Information Interchange
HTML	Hypertext Markup Language
OSI	Open Systems Interconnection
SGML	Standard Generalized Markup Language
URL	Uniform Resource Locator
UCS	Universal Multiple-Octet Coded Character Set

4 Layer structure

4.1 The three layer stack

The general model described by this Technical Report represents each real system of the communcation process as a three layer stack composed of

a) a User Transformation Layer;

b) an Application Transformation Layer;

c) an Interchange Transformation Layer.

This is illustrated in figure 1, which also shows the underlying binary transport service and the peer-level transformations described in clause 5 below. All layers are permitted to have an internal sub-layer structure which represents their overall transformation as the result of a sequence of separately specified transformations.

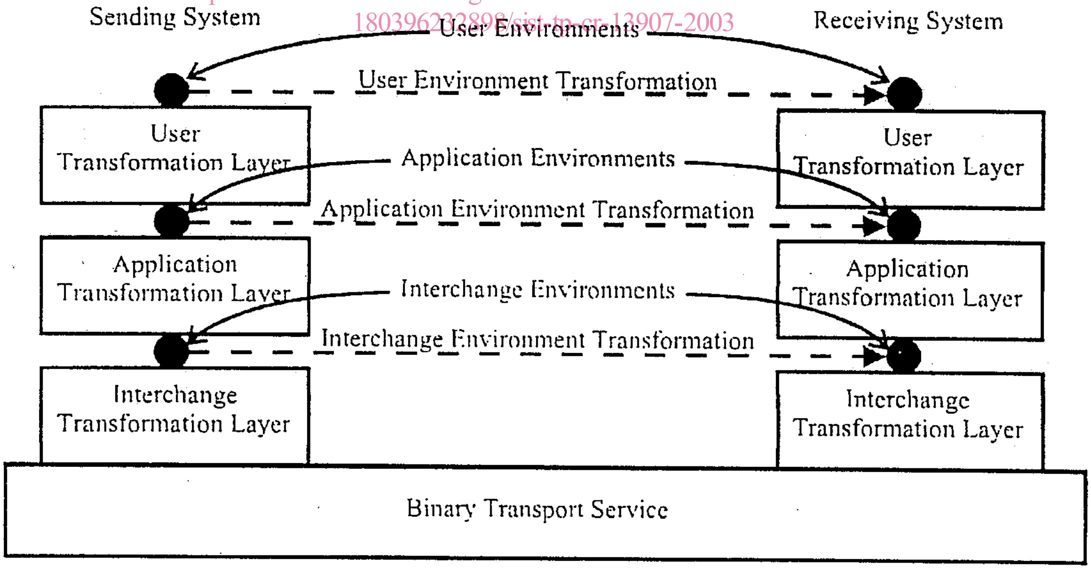

Figure 1 — Layer structure of the transformation model

Character data is conceptually passed through the stack of the sending system from top to bottom, emerging as a byte stream which is transported transparently to the receiving system, where it is passed through the receiving stack from bottom to top to emerge once again as character data.

A single real system normally possesses devices both for input and output and has a bidirectional communications system. Such a system is modelled with common application and interchange transformation layers for sending and receiving but with a separate user transformation layer for each device. The user transformation layers will share a common application environment, as illustrated in figure 2. In modelling a particular real system there may be ambiguity as to whether a sublayer representing some aspect of the system should be considered as part of the user or the application transformation layer. If the system is bidirectional, this can be resolved, with reference to figure 2, by considering whether the sublayer is invoked in both, or only one of, sending and receiving.

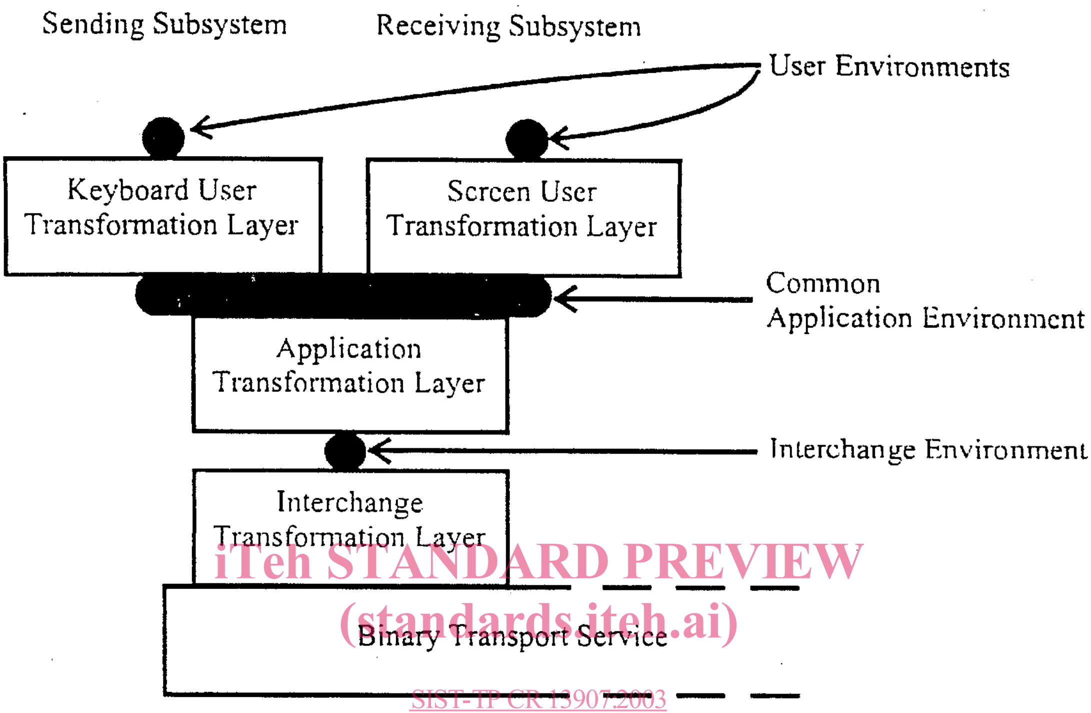
Sending Subsystem
Receiving Subsystem
User Environments
Keyboard User Transformation Layer
Screen User Transformation Layer
Common Application Environment
Application Transformation Layer
Interchange Environment
Interchange Transformation Layer
Binary Transport Service

Figure 2 — Layer structure of a bidirectional system

4.2 Nature of binary transport

The processes involved in the binary transport service may involve compression and encryption, unreliable links and associated error recovery processes, and any other processes which are performed without regard to the interpretation of the bytes being transported. All such processes are outside the scope of this report, as are means of addressing within any network that may be involved. If any intermediate system involved in the transport process, however, performs operations that depend on the interpretation of the byte stream as characters then that intermediate system should be considered as a relay system composed conceptually of two linked end systems, each of which falls within the scope of the model. The link between these end systems is through a common Application Environment which passes characters transparently between the two Application Transformation Layers as illustrated in figure 3. The User Transformation Layers are absent in such a relay. The content dependent processing that occurs in the relay is then represented within the model through the sending and receiving stacks of the model operating according to different standards. In particular the interchange environments of the two systems comprising the relay will generally differ, in contrast to those of two systems communicating as in figure 1 where the interchange environments are generally identical.