TECHNICAL REPORT

ISO/IEC TR 15285

First edition
1998-12-15

Information technology — An operational model for characters and glyphs

Technologies de l'information — Modèle pour l'utilisation de caractères graphiques et de glyphes

Reference number
ISO/IEC TR 15285:1998(E)

Contents

ISO/IEC Copyright Office • Case Postale 56 • CH-1211 Genève 20 • Switzerland

Printed in Switzerland

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 committate in fields of mutual interest.
Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

The main task of a technical committee is to prepare International Standards, but in exceptional circumstances a technical committee may propose the publication of a Technical Report of one of the following types:

— Type 1, when the required support cannot be obtained for the publication of an International Standard, despite repeated efforts;

— Type 2, when the subject is still under technical development or where for any other reason there is the future but not immediate possibility of an agreement on an International Standard;

— Type 3, when a technical committee has collected data of a different kind from that which is normally published as an International Standard ("state of the art", for example).

Technical Reports of types 1 and 2 are subject to review within three years of publication to decide whether they can be transformed into International Standards. Technical Reports of type 3 do not necessarily have to be reviewed until the data they provide are considered to be no longer valid or useful.

ISO/IEC TR 15285, which is a Technical Report of type 3, was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 2, *Coded character sets*, and Subcommittee SC 18, *Document processing and related communication* (which has since been reorganized into SC 34, Document description and processing languages).

Introduction

People interpret the meaning of a written sentence by the shapes of the characters contained in it. For the characters themselves, people consider the information content of a character inseparable from its printed image. Information technology, in contrast, makes a distinction between the concepts of a character's meaning (the information content) and its shape (the presentation image). Information technology uses the term *character* (or *coded character*) for the information content, and the term *glyph* for the presentation image. A conflict exists because people consider characters and glyphs equivalent. Moreover, this conflict has led to misunderstanding and confusion. This Technical Report provides a framework for relating characters and glyphs to resolve the conflict because successful processing and printing of character information on computers requires an understanding of the appropriate use of characters and glyphs.

Historically, ISO/IEC JTC 1/SC 2 has had responsibility for the development of coded character set standards such as ISO/IEC 10646 for the digital representation of letters, ideographs, digits, symbols, etc. ISO/IEC JTC 1/SC 18 has had responsibility for the development of standards for document processing, which presents the characters coded by SC 2. SC 18 standards include the font standard, ISO/IEC 9541, and the glyph registration standard, ISO/IEC 10036. The Association for Font Information Interchange (AFII) maintains the 10036 glyph registry on behalf of ISO.

This Technical Report is written for a reader who is familiar with the work of SC 2 and SC 18. Readers without this background should first read Annex B, "Characters", and Annex C, "Glyphs".

This edition of the Technical Report does not fully develop the complex issues associated with the Chinese, Japanese, Korean, and Vietnamese ideographic characters used in East Asia. In addition, although it discusses the process of rendering digital character information for display and printing, it avoids discussing the inverse process of character recognition (that is, converting printed text into character information in the computer).

Information technology — An operational model for characters and glyphs

1 Scope

The purpose of this Technical Report is to provide a general framework for discussing characters and glyphs. The framework is applicable to a variety of coded character sets and glyph-identification schemes. For illustration, this Technical Report uses examples from characters coded in ISO/IEC 10646 and glyphs registered according to ISO/IEC 10036.

This Technical Report

— differentiates between coded characters and registered glyphs

— identifies the domain of use of coded characters and glyph identifiers

— provides a conceptual framework for the formatting and presentation of coded character data using glyph identifiers and glyph representations

This Technical Report describes idealized principles that were not completely followed in coding characters for ISO/IEC 10646 and in registering glyphs according to ISO/IEC 10036. The fact that ISO/IEC 10646, ISO/IEC 10036, and other standards do not completely follow the principles in the model does not invalidate the model and does not diminish the utility of having the model.

2 References

ISO/IEC 9541-1: 1991, *Information technology — Font information interchange — Part 1: Architecture.*

ISO/IEC 10036: 1996, *Information technology — Font information interchange — Procedures for registration of font-related identifiers.*

ISO/IEC 10180: 1995, *Information technology — Processing languages — Standard Page Description Language (SPDL).*

ISO/IEC 10646-1: 1993, *Information technology — Universal Multiple-Octet Coded Character Set (UCS) — Part 1: Architecture and Basic Multilingual Plane.*

3 Definitions

For the purpose of this Technical Report, the following definitions apply. The definitions have been extracted from the ISO/IEC 9541-1: 1991 and ISO/IEC 10646-1: 1993 standards.

3.1 character: A member of a set of elements used for the organisation, control, or representation of data. (ISO/IEC 10646-1: 1993)

3.2 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. (ISO/IEC 10646-1: 1993)

3.3 font: A collection of glyph images having the same basic design, e.g. Courier Bold Oblique. (ISO/IEC 9541-1: 1991)

3.4 font resource: A collection of glyph representations together with descriptive and font metric information which are relevant to the collection of glyph representations as a whole. (ISO/IEC 9541-1: 1991)

3.5 glyph: A recognizable abstract graphic symbol which is independent of any specific design. (ISO/IEC 9541-1: 1991)

3.6 glyph collection: An identified set of glyphs. (ISO/IEC 9541-1: 1991)

3.7 glyph image: An image of a glyph, as obtained from a glyph representation displayed on a presentation surface. (ISO/IEC 9541-1: 1991) [See the definition of *graphic symbol*.]

3.8 glyph metrics: The set of information in a glyph representation used for defining

the dimensions and positioning of the glyph shape. (ISO/IEC 9541-1: 1991)

3.9 glyph representation: The glyph shape and glyph metrics associated with a specific glyph in a font resource. (ISO/IEC 9541-1: 1991)

3.10 glyph shape: The set of information in a glyph representation used for defining the shape which represents the glyph. (ISO/IEC 9541-1: 1991)

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

3.12 graphic symbol: The visual representation of a graphic character or of a composite sequence. (ISO/IEC 10646-1: 1993) [See the definition of *glyph image*.]

3.13 presentation [of a graphic symbol]: The process of writing, printing, or displaying a graphic symbol. (ISO/IEC 10646-1: 1993)

3.14 presentation form: In the presentation of some scripts, a form of a graphic symbol representing a character that depends on the position of the character relative to other characters. (ISO/IEC 10646-1: 1993)

3.15 presentation surface: A virtual representation of a presentation medium (page, graphic display, etc.) maintained by the presentation process, on which all glyph shapes are to be imaged. (ISO/IEC 9541-1: 1991)

3.16 repertoire: A specified set of characters that are represented in a coded character set. (ISO/IEC 10646-1: 1993)

4 Character and glyph distinctions

The character and glyph definitions in clause 3, which were taken from ISO/IEC 10646 and ISO/IEC 9541, were developed independently and contain terminology that requires explanation.

In information technology, *characters* are abstract information elements in the domain of coding for data representation and, in particular, data interchange. Coded character set standards assign numeric values, character names, and representative (sample) images to each character contained in a coded character set. Typically a character is given a name, which also serves to differentiate it from the other characters of the coded character set. The precise semantics and appearance of the information elements in any given implementation are not defined by those standards for coded character sets. This apparent lack of definition is not considered to be a defect in the standards. Recognizing that the information may be acted upon (deciphered, sorted, transformed, formatted, archived, presented, etc.) by many different application processes during its lifetime, standards for coded character sets are defined as a basis for information interchange.

In information technology, *glyphs* are abstract presentation elements in the domain of presentation processing. The ISO/IEC 10036 standard for glyph registration defines the process for assigning glyph identifiers, glyph descriptions, and representative (sample) images to each glyph submitted for registration. The precise usage and appearance of these presentation elements in any implemented font resource is not defined by those glyph registration activities. As with the coded character set standards, this apparent lack of definition is not considered to be a defect in the standards. Glyph identifiers are unambiguously assigned as a basis for tagging presentation elements in and among interchanged font resources, recognizing that the font-specific design information may vary from one font resource to another.

Characters and *glyphs* are closely related, with many attributes in common and yet with distinctions that make it essential that they be managed in information processing as separate entities. The ISO/IEC 10646 standard recognizes the distinction between

characters and their visual representation by defining the term, *graphic symbol*. The *graphic symbol* of SC 2 standards and the *glyph image* of SC 18 standards represent equivalent concepts. However, *glyph* and its associated ISO/IEC 9541 terminology are preferred when referring to presentation and presentation processing.

The historical association of characters and glyphs has resulted in character sets maintaining distinctions that cannot be founded on distinctions in meaning, but only on distinctions in shape. Similarly, the glyph registration authority and the SC 18 font resource model have made use of criteria based on meaning to abstract potential distinctions in shape. In practice, ISO/IEC 10646 contains characters that appear to be instances of glyphs, while the glyph registry prescribed by ISO/IEC 10036 contains glyphs that appear to be designated as abstract characters. In both cases, the ideal nature of characters and glyphs has been compromised to a degree. For example, in ISO/IEC 10646-1, SC 2 coded the "fi" glyph into the character U+FB01 LATIN SMALL LIGATURE FI "fi" character, and the U+0410 CYRILLIC CAPITAL LETTER A "A" character. However, AFII instead registered three glyph identifiers.

Within the realm of information technology, an ideal characterization of characters and glyphs and their relationship may be stated as follows:

— A *character* conveys distinctions in meaning or sounds. A character has no intrinsic appearance.

1) This Technical Report describes a character in terms of its 10646 code position (U+FB01), its 10646 name (LATIN SMALL LIGATURE FI), and illustrates it with a representative glyph in quotation marks ("

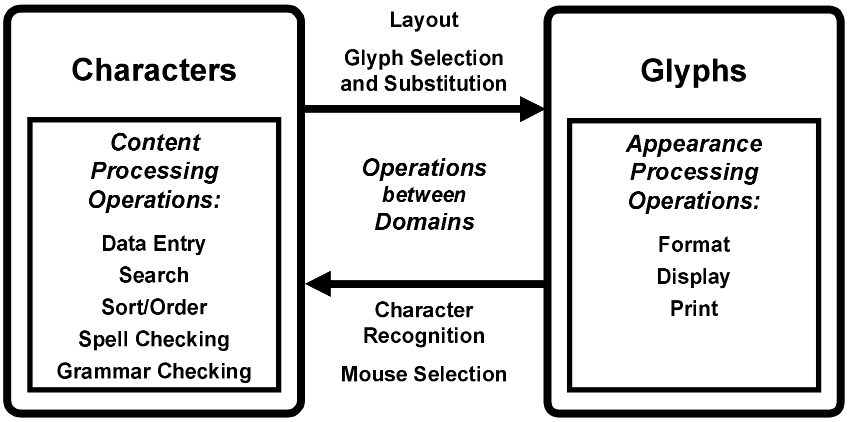

Figure 1 — Character and glyph domains

ages into coded characters. Also, a paragraph-level hyphenation process is an example of a layout process that requires content information.

It is not possible, in general, to code data in such a way as to optimize one process without reducing the performance of other processes. Even within the content domain, the nature of the character coding employed for textual data affects the type or types of processing to be performed on the data; no single coding can optimize more than a few such potential processes. Given this situation, the best solution is to formulate an independent, logical character coding that, when necessary, can be transformed into another coding more amenable to the processing required. For example, in the case of searching, character data is often recast into specific forms that facilitate quick searches. For sorting, a specially created sort key is required. In addition, because ISO/IEC 10646 contains glyph-like characters, it is expected that implementations may choose to canonicalize or normalize such characters by translating them to normative characters. A presentation subsystem that employs such a technique may require that character data be normalized prior to presentation.

The recognition that two separate domains of processing are commonly applied to character-based information leads to a conclusion that two primary forms of this information are needed:

1. a content-oriented form that is amenable to immediate content-based processes and that can be easily converted to and from other optimized forms
2. an appearance-oriented form that facilitates imaging of content

These are, respectively, the character-based form and the glyph-based form. Failure to recognize this distinction between the character domain and the glyph domain has led to the development of inconsistent standards and inconsistent systems that lack functional separation of the two domains.

5.2 Composition, layout, and presentation

As depicted in Figure 2 on the next page, the composition and layout process (for glyph selection and positioning) spans both processing domains. If attention is restricted to the text portion of this process, the presentation of character-based information requires three primary operations:

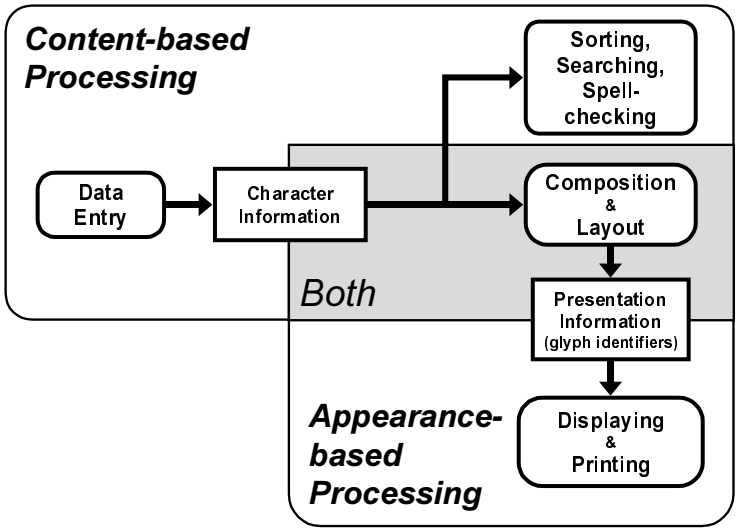

Figure 2 — Composition, layout, and presentation

— selecting the glyph representations needed to display character data

— positioning the glyph shapes on the presentation surface

— imaging the glyph shapes

Glyph selection is the process of selecting (possibly through several iterations) the most appropriate glyph identifier or combination of glyph identifiers to render a coded character or composite sequence of coded characters. Coded characters and their associated implicit or explicit formatting information (for example, specification of the font and its size) represent the primary inputs to composition and layout processing, and glyph identifiers (or the associated glyph metrics and glyph shapes) represent the primary output from composition and layout processing. The degree of glyph selection sophistication varies widely among existing standards and implementations.

The relationship between coded characters and glyph identifiers may be one-to-one, one-to-many, many-to-one, or many-to-many.[3] This is particularly true for ISO/IEC 10646 implementation level 3, which uses combining characters. In its fully general form, the relationship is a context-sensitive M-to-N mapping where $M > 0$, $N \geq 0$. For some characters in ISO/IEC 10646-1, for example, the U+FEFF ZERO WIDTH NO-BREAK SPACE character, no glyph (N=0) is defined.

The SC 18 document-processing model separates the glyph selection and layout operations from the operation of imaging the glyph shape to permit document interchange between the processes. Glyph selection and positioning are part of the composition and layout process, whereas imaging the glyph shape is part of the presentation process. The result of composition and layout is a *final-form document*, which contains font identifiers, glyph identifiers, and coordinate positions, along with either references to font resources or the actual font resources themselves. Such a document form contains all the necessary information required to present the formatted

3) The necessity for mapping characters to glyphs (glyph selection), not its complexity, is one of the motivations for developing this operational model for characters and glyphs.

document on some presentation medium. An example of such a final form document is an SPDL (ISO/IEC 10180) document instance.

An important aspect of this document-processing model is that it begins with coded-character data as its input and produces either glyph-based data or directly imaged glyph shapes as its output. That is, it incorporates a transformation from a coded-character representation of a document's *content* to a glyph-based coding of a document's *appearance*. The latter may only be visible to the internal mechanisms of an operating system or a user-interface subsystem in the case that the result is directly imaged for presentation. However, even these systems frequently support some form of output that contains the glyph-based final form of the document.

6 Glyph selection

While some earlier formatting systems assume a one-to-one correspondence between characters and glyphs, this is inadequate for many applications and scripts. Many contemporary composition and layout systems support more complex glyph-selection processes that provide for the representation of sequences of multiple character codes by a single glyph or by the use of sequences of glyphs to represent certain characters. In general, glyph selection needs to be based on style information and context as well as on the character data itself. For example, consider the following:

— When the U+0022 QUOTATION MARK "" character is encountered, a composition and layout process may need to determine whether it begins or ends a quotation and then choose either an opening or closing quotation mark glyph ("“" or "”") as appropriate. In addition, the process may select glyphs depending on the language of the text being formatted (or the formatting style specifications that apply to the content being formatted). For example, German text could use the "„" and "“" glyphs for quotation marks; and French text, the "«" and "»" glyphs.

— When the U+002D HYPHEN-MINUS "–" character is encountered, a composition and layout process may have to determine if it is used in a mathematical formula, as a separator between figures (digits), as a separator between words, or as a separator between syllables. Depending on which context applies, it will select a minus sign, a figure dash, a quotation dash, or a hyphen dash (or possibly a hyphen point) glyph to display the character.

NOTE: Because the ISO/IEC 10646 repertoire includes the necessary characters, some applications resolve quotation marks and the hyphen-minus illustrated in the previous two points by converting to the appropriate 10646 *characters* as they are input rather than selecting the appropriate glyphs for presentation.

— When a parenthesis or square bracket character is encountered in a document being formatted in vertical lines (for example, with East Asian ideographs), a composition and layout process may need to choose a vertical variant glyph form of the parenthesis or square bracket. It may also perform a similar selection for certain other characters such as U+30FC KATAKANA-HIRAGANA PROLONGED SOUND MARK "ー"

" are shown in Figure 3 on the next page.

— In addition, Arabic typography makes extensive use of ligatures. For example, Figure 4 shows the isolated forms of U+0627 ARABIC LETTER ALEF "ا", and then the two ligature forms used when Lam is followed by Alef.

— When a U+0930 DEVANAGARI LETTER RA "र" is encountered in a Hindi, Marathi, Sanskrit, etc. document, a composition and layout process may have to determine whether a subscript, superscript, half ("eyelash"), or full form glyph is required according to context. If a subscript form is required, a composition and layout process may have to choose from one of a number of possible subscript forms depending on the glyph to which it is to be attached. Figure 5 shows an example of this.

The process of glyph selection is sometimes implemented as a separate part of composition and layout because many of the choices required to determine an appropriate glyph are based solely on (1) the context of a character within a document, (2) the style specifications that apply to a given character, or (3) a combination of the context and style specification. All of the choices required for the examples shown above fall into one of these categories. However, in general, glyph selection can only be made as an integral part of the entire composition and layout process. Consider the following:

— When hyphenating a line of text during composition, a composition and layout process may insert a hyphen glyph form at the end of a line if the line is broken at a hyphenation point.

— If hyphenating a German text between the letters "c" and "k", a composition and layout process may replace the "c" with a "k".

— If during the composition of a German text, the character sequence "ff " is encountered, a composition and layout process may select two distinct (non-ligated) glyph forms for U+0066 LATIN SMALL LETTER F "f". However, if the position for a hyphen (a hyphen point) should occur before the last "f", that is, at "ff-f", then a composition and layout process may select an ff ligature glyph "ff", followed by a hyphen (on the first line), and begin the subsequent line with a normal glyph for the third and final "f".

— A composition and layout process may select small cap glyph forms for the first line of a paragraph of Roman text.

— A composition and layout process may select a swash glyph form for the first and last character of each line of a paragraph.

— A composition and layout process may select one of a number of possible variant glyph forms for certain Arabic letters depending on whether more or less space is available for composing a line of Arabic text.

— When justifying a line of Arabic text, a composition and layout process may start by selecting ligature glyph forms that consume the smallest amount of linear space in a line, and then sequentially replace these ligatures with component ligatures or component non-ligature glyphs such that more linear line space is consumed up to the required line measure. Alternatively, a composition and layout process may start justification by selecting no ligatures and then sequentially select ligatures that consume a smaller amount of linear space until the desired line measure is achieved or until an inter-word space stretch threshold is reached (that is, a point at which inter-word spaces can be stretched to justify the line to the desired measure).

In summary, the glyph-selection process is primarily applicable to behavior occurring at the end or beginning of individual lines of text, or within the context of justifying or altering the measure of a given line during line composition. A system supporting the capabilities illustrated in the preceding examples must include glyph selection as an integral part of the composition and layout process.

7 Summary

Here are the primary points of this technical report:

— Most people equate a character and its shape.

— This causes difficulties and misunderstanding because contemporary information technology distinguishes two related, but distinct, domains:

  – The processing domain uses coded characters to represent the character's meaning.
  – The presentation domain uses glyph identifiers to represent the character's image.

— Processes are available to convert between the two domains:

  – Presentation processing takes the coded-character data plus any formatting data plus font information to display and print character data.
  – A character recognition process scans images, analyzes the shapes, and outputs the coded characters that correspond to the shapes.

— Depending on the script and the particular font or fonts used, glyph selection can be straightforward or relatively complex.

  – It is straightforward when a one-to-one correspondence exists between the set of coded characters and the set of glyphs in a font.
  – The process is more complex when it must choose between several alternatives; for example, when a sequence of coded characters may be mapped into more than one sequence of glyphs in a font.