Compound Text Encoding

X Consortium Standard

Robert W. Scheifler, X Consortium

X Version 11, Release 7.7

Version 1.1

Copyright © 1989 X Consortium

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED “AS IS”, WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE X CONSORTIUM BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

Except as contained in this notice, the name of the X Consortium shall not be used in advertising or otherwise to promote the sale, use or other dealings in this Software without prior written authorization from the X Consortium.

X Window System is a trademark of The Open Group.

Table of Contents

Overview .....
Values .....
Control Characters .....
Standard Character Set Encodings .....
Approved Standard Encodings .....
Non-Standard Character Set Encodings .....
Directionality .....
Resources .....
Font Names .....
Extensions .....
Errors .....

Overview

Compound Text is a format for multiple character set data, such as multi-lingual text. The format is based on ISO standards for encoding and combining character sets. Compound Text is intended to be used in three main contexts: inter-client communication using selections, as defined in the (ICCCM) window properties (e.g., window manager hints as defined in the ICCCM) and resources (e.g., as defined in Xlib and the Xt Intrinsics).

Compound Text is intended as an external representation, or interchange format, not as an internal representation. It is expected (but not re uired) that clients will convert Compound Text to some internal representation for processing and rendering, and convert from that internal representation to Compound Text when providing textual data to another client.

Values

The name of this encoding is "COMPOUND TEXT". When text values are used in the ICCCM-compliant selection mechanism or are stored as window properties in the server, the type used should be the atom for "COMPOUND TEXT".

Octet values are represented in this document as two decimal numbers in the form col/row. This means the value (col 1) row. For example, / 1 means the value .

For our purposes, the octet encoding space is divided into four ranges:

C	octets from	/	to 1/1
GL	octets from	/	to /1
C1	octets from	/	to /1
GR	octets from 1	/	to 1 /1

C and C1 are "control character" sets, while GL and GR are "graphic character" sets. Only a subset of C and C1 octets are used in the encoding, and depending on the character set encoding defined as GL or GR, a subset of GL and GR octets may be used see below for details. All octets (/ to 1 /1) may appear inside the text of extended segments (defined below).

For those familiar with ISO , we will use only an -bit environment, and we will always use G for GL and G1 for GR.

Control Characters

In C , only the following values will be used:

/	HT	HORI ONTAL TABULATION
/1	NL	NEW LINE
1/11	ESC	(ESCAPE)

In C1, only the following value will be used:

/11	CSI	CONTROL SE UENCE INTRODUCER

The alternate -bit CSI encoding 1/11 /11 is not used in Compound Text.

No control se uences are defined in Compound Text for changing the C and C1 sets.

A horizontal tab can be represented with the octet / . Specification of tabulation width settings is not part of Compound Text and must be obtained from context (in an unspecified manner).

Inclusion of horizontal tab is for consistency with the STRING type currently defined in the ICCCM.

A newline (line separator/terminator) can be represented with the octet /1 .

Note that /1 is normally LINEFEED, but is being interpreted as NEWLINE. This can be thought of as using the (deprecated) NEW LINE mode, E.1. , in ISO . Use of this value instead of / (NEL, NEXT LINE) is for consistency with the STRING type currently defined in the ICCCM.

The remaining C and C1 values (1/11 and /11) are only used in the control se uences defined below.

Standard Character Set Encodings

The default GL and GR sets in Compound Text correspond to the left and right halves of ISO -1 (Latin 1). As such, any legal instance of a STRING type (as defined in the ICCCM) is also a legal instance of type COMPOUND TEXT.

The implied initial state in ISO is defined with the se uence: 1/11 / / GO and G1 in an -bit environment only. Designation also invokes. 1/11 / / In an -bit environment, C1 represented as -bits. 1/11 / / Graphic character sets can be or . 1/11 / /11 -bit code is used. 1/11 / / Designate ASCII into G . 1/11 /1 / 1 Designate right-hand part of ISO Latin-1 into G1.

To define one of the approved standard character set encodings to be the GL set, one of the following control se uences is used:

1/11 / I F character set
1/11 / / I F character set

To define one of the approved standard character set encodings to be the GR set, one of the following control se uences is used:

1/11 / I F character set
1/11 /1 I F character set
1/11 / / I F character set

The "F"in the control se uences above stands for "Final character", which is always in the range / to /1 . The " I " stands for zero or more "intermediate characters", which are always in the range / to /1 , with the first intermediate character always in the range / 1 to / . The registration authority has defined an " I F" se uence for each registered character set encoding.

Final characters for private encodings (in the range / to /1) are not permitted here in Compound Text.

For GL, octet / is always defined as SPACE, and octet /1 (normally DELETE) is never used. For a -character set defined as GR, octets 1 / and 1 /1 are never used.

This is consistent with ISO .

A N character set uses N octets (N 1) for each character. The value of N is derived from the column value for F:

column	or	octets
column		octets
column		or more octets

In a N encoding, the octet values / and /1 (in GL) and 1 / and 1 /1 (in GR) are never used.

The column definitions come from ISO .

Once a GL or GR set has been defined, all further octets in that range (except within control se uences and extended segments) are interpreted with respect to that character set encoding, until the GL or GR set is redefined. GL and GR sets can be defined independently, they do not have to be defined in pairs.

Note that when actually using a character set encoding as the GR set, you must force the most significant bit (/) of each octet to be a one, so that it falls in the range 1 / to 1 /1 .

Control se uences to specify character set encoding revisions (as in section . .1 of ISO) are not used in Compound Text. Revision indicators do not appear to provide useful information in the context of Compound Text. The most recent revision can always be assumed, since revisions are upward compatible.

Approved Standard Encodings

The following are the approved standard encodings to be used with Compound Text. Note that none have Intermediate characters however, a good parser will still deal with Intermediate characters in the event that additional encodings are later added to this list.

/	-bit ASCII graphics (ANSI X . -1), Left half of ISO sets
/	Right half of IS X 1-1 (rea irmed 1), -Bit Alphanumeric-Katakana Code
/1	Left half of IS X 1-1 (rea irmed 1), -Bit Alphanumeric-Katakana Code
/ 1	Right half of ISO -1, Latin alphabet No. 1
/	Right half of ISO - , Latin alphabet No.
/	Right half of ISO - , Latin alphabet No.
/	Right half of ISO - , Latin alphabet No.

/	Right half of ISO - , Latin/Greek alphabet
/	Right half of ISO - , Latin/Arabic alphabet
/	Right half of ISO - , Latin/Hebrew alphabet
/1	Right half of ISO - , Latin/Cyrillic alphabet
/1	Right half of ISO - , Latin alphabet No.
/ 1	GB 1 -1 , China (PRC) Hanzi
/	IS X -1 , apanese Graphic Character Set
/	KS C 1-1 , Korean Graphic Character Set

The sets listed as "Left half of ..." should always be defined as GL. The sets listed as "Right half of ..." should always be defined as GR. Other sets can be defined either as GL or GR.

Non-Standard Character Set Encodings

Character set encodings that are not in the list of approved standard encodings can be included using "extended segments". An extended segment begins with one of the following se uences:

1/11 / /1 / M variable number of octets per character
L

1/11 / /1 / 1 M 1 octet per character
L

1/11 / /1 / M octet per character
L

1/11 / /1 / M octet per character
L

1/11 / /1 / M octet per character
L

This uses the "other coding system" of ISO , using private Final characters.

The "M" and "L" octets represent a 1 -bit unsigned value giving the number of octets that appear in the remainder of the segment. The number is computed as ((M - 1) 1) (L - 1). The most significant bit M and L are always set to one. The remainder of the segment consists of two parts, the name of the character set encoding and the actual text. The name of the encoding comes first and is separated from the text by the octet / (STX, START OF TEXT). Note that the length defined by M and L includes the encoding name and separator.

The encoding of the length is chosen to avoid having zero octets in Compound Text when possible, because embedded NUL values are problematic in many C language routines. The use of zero octets cannot be ruled out entirely however, since some octets in the actual text of the extended segment may have to be zero.

The name of the encoding should be registered with the X Consortium to avoid con icts and should when appropriate match the CharSet Registry and Encoding registration used in the X Logical Font Description. The name itself should be

encoded using ISO -1 (Latin 1), should not use uestion mark (/1) or asterisk (/1), and should use hyphen (/1) only in accordance with the X Logical Font Description.

Extended segments are not to be used for any character set encoding that can be constructed from a GL/GR pair of approved standard encodings. For example, it is incorrect to use an extended segment for any of the ISO family of encodings.

It should be noted that the contents of an extended segment are arbitrary for example, they may contain octets in the C and C1 ranges, including / , and octets comprising a given character may di er in their most significant bit.

ISO-registered "other coding systems" are not used in Compound Text extended segments are the only mechanism for non- encodings.

Directionality

If desired, horizontal text direction can be indicated using the following control se uences:

/11	/ 1	/1	begin left-to-right text
/11	/	/1	begin right-to-left text
/11	/1		end of string

This is a subset of the SDS (START DIRECTED STRING) control in the Draft Bidirectional Addendum to ISO .

Directionality can be nested. Logically, a stack of directions is maintained. Each of the first two control se uences pushes a new direction on the stack, and the third se uence (revert) pops a direction from the stack. The stack starts out empty at the beginning of a Compound Text string. When the stack is empty, the directionality of the text is unspecified.

Directionality applies to all subse uent text, whether in GL, GR, or an extended segment. If the desired directionality of GL, GR, or extended segments di ers, then directionality control se uences must be inserted when switching between them.

Note that definition of GL and GR sets is independent of directionality defining a new GL or GR set does not change the current directionality, and pushing or popping a directionality does not change the current GL and GR definitions.

Specification of directionality is entirely optional text direction should be clear from context in most cases. However, it must be the case that either all characters in a Compound Text string have explicitly specified direction or that all characters have unspecified direction. That is, if directionality control se uences are used, the first such control se uence must precede the first graphic character in a Compound Text string, and graphic characters are not permitted whenever the directionality stack is empty.

Resources

To use Compound Text in a resource, you can simply treat all octets as if they were ASCII/Latin-1 and just replace all " " octets (/1) with the two octets " ",

all newline octets (/1) with the two octets " n", and all zero octets with the four octets " ". It is up to the client making use of the resource to interpret the data as Compound Text the policy by which this is ascertained is not constrained by the Compound Text specification.

Font Names

The following CharSet names for the standard character set encodings are registered for use in font names under the X Logical Font Description:

ISO -1	ISO -1	Latinalphabet No. 1
ISO -	ISO -	Latinalphabet No.
ISO -	ISO -	Latinalphabet No.
ISO -	ISO -	Latinalphabet No.
ISO -	ISO -	Latin/Cyrillic alphabet
ISO -	ISO -	Latin/Arabic alphabet
ISO -	ISO -	Latin/Greekalphabet
ISO -	ISO -	Latin/Hebrew alphabet
ISO -	ISO -	Latinalphabet No.
ISX 1.1 - 1)	IS X 1-1 (rea irnd	Alphanumeric-Katakana Code
GB 1 .1 -	GB 1 -1 , GL encoding	China (PRC) Hanzi
ISX .1 -	IS X -1 , GL encoding	apanese Graphic Charac
KSC 1.1 -	KS C 1-1 , GL encoding	Korean Graphic Character Set

Extensions

There is no absolute re uirement for a parser to deal with anything but the particular encoding syntax defined in this specification. However, it is possible that Compound Text may be extended in the future, and as such it may be desirable to construct the parser to handle / syntax more generally.

There are two general formats covering all control se uences that are expected to appear in extensions:

1/11 I F

For this format, I is always in the range / to /1 , and F is always in the range / to /1 .

/11 P I F

For this format, P is always in the range / to /1 , I is always in the range / to /1 , and F is always in the range / to /1 .

In addition, new (singleton) control characters (in the C and C1 ranges) might be defined in the future.

Finally, new kinds of "segments" might be defined in the future using syntax similar to extended segments:

1/11 / /1 F M L

For this format, F is in the range / to /1 . M and L are as defined in extended segments. Such a segment will always be followed by the number of octets defined by M and L. These octets can have arbitrary values and need not follow the internal structure defined for current extended segments.

If extensions to this specification are defined in the future, then any string incorporating instances of such extensions must start with one of the following control se uences:

1/11 / V /	ignoring extensions is OK
1/11 / V / 1	ignoring extensions is not OK

In either case, V is in the range / to /1 and indicates the major version minus one of the specification being used. These version control se uences are for use by clients that implement earlier versions, but have implemented a general parser. The first control se uence indicates that it is acceptable to ignore all extension control se uences no mandatory information will be lost in the process. The second control se uence indicates that it is unacceptable to ignore any extension control se uences mandatory information would be lost in the process. In general, it will be up to the client generating the Compound Text to decide which control se uence to use.

Errors

If a Compound Text string does not match the specification here (e.g., uses undefined control characters, or undefined control se uences, or incorrectly formatted extended segments), it is best to treat the entire string as invalid, except as indicated by a version control se uence.