Stable Implementation Agreements for Open Systems Interconnection Protocols: Part 21 - CHARACTER SET

Output from the December 1993 Open Systems Environment Implementors' Workshop (OIW)

SIG Chair: **Wally Wedel, U. S. West Advanced Technologies**
Workshop Editor: Brenda Gray, NIST

Foreword

This part of the Working Implementation Agreements was prepared by the Character Set Working Group, formerly affiliated with the Upper Layer Special Interest Group of the National Institute of Standards and Technology (NIST) Workshop for Implementors of Open Systems Interconnection (OSI). See Part 1 - General Information of the "Draft Working Implementors' Agreements Document" for the workshop charter. Text in this part has been approved by the Plenary of the above-named workshop.

Table of Contents

List of Figures

List of Tables

Part 21 - Character Set Usage in OSI Applications

This International Standardized Profile is defined within the context of Functional Standardization, in accordance with the principles specified by ISO TR 10000, "Taxonomy Framework and Directory of Profiles." The context of Functional Standardization is one part of the overall field of Information Technology (IT) standardization activities, covering base standards, profiles, and registration mechanisms. A profile defines a combination of base standards that collectively perform a specific well-defined IT function. Profiles standardize the use of options and other variations in the base standards, and provide a basis for the development of uniform, internationally recognized system tests.

This International Standardized Profile was developed in close cooperation between the three International OSI Workshops: the NIST OSI Implementors Workshop (NIST OIW), the European Workshop for Open Systems (EWOS), and the Asia Oceanic Workshop (AOW). The text is harmonized between these three Workshops and was ratified by the Workshops' plenary assemblies.

This International Standardized Profile contains an informative Annex A - Character Set Technology.

1 Scope

This International Standardized Profile describes Information Processing Character Set agreements covering character set usage in referencing Application Service Elements and OSI Applications. These agreements are based upon ISO Character Set International Standards and CCITT Character Set Recommendations. The informative Annex A summarizes the character set practices within referencing Application Service Elements and OSI Applications including all relevant encoding information drawn from the appropriate ISO Registers, ISO Standards, and CCITT Recommendations.

1.1 Recording Additional Character Sets

This International Standardized Profile does not prevent Application Service Elements from adding new graphic character sets or control function sets. When new character sets are to be added, however, they shall be recorded in the International Standardized Profile.

1.2 General Assembly of Character Sets

For the purpose of this International Standardized Profile when new character sets are to be added, efforts shall be made to obtain agreement on their uses among Application Service Elements so that they are generally applicable.

1.3 Minimum Number of Character Sets

The number of character sets supported will be kept to the minimum possible so as to maximize interoperability.

2 Normative References

The following International Standards and CCITT Recommendations are referenced in this International Standardized Profile:

2.1 CCITT

[**1**] CCITT Recommendations T.100, 1985, *International Information Exchange for Videotex.*

[**2**] CCITT Recommendation T.50, 1985, *International Alphabet No. 5.*

[**3**] CCITT Recommendation T.51, 1985, *Coded Character Sets for Telematic Services.*

[**4**] CCITT Recommendation T.61, 1985, *Character Repertoire and Coded Character Sets for the International Teletex Service.*

2.2 ISO

[**5**] DIS 8859-7, 1987, *Information Processing - 8 bit single-byte coded graphic character sets - Part 7: Latin/Greek alphabet.*

[**6**] IS 2022, 1986, *Information Processing - ISO 7-bit and 8-bit coded character sets - Code extension techniques.*

[**7**] IS 2375, 1985, *Data Processing - Procedure for registration of escape sequences.*

[**8**] IS 4873, 1986, *Information Processing - ISO 8-bit code for information interchange - Structure and rules for implementation.*

[**9**] IS 6429, 1983, *Information Processing - ISO 7-bit and 8-bit coded character sets - Additional control functions for character-imaging devices.*

[**10**] IS 646, 1983, *Information Processing - ISO 7-bit coded character set for information interchange.*

[**11**] IS 6937/1, 1983, *Information Processing - Coded character sets for text communication - Part 1: General introduction.*

[**12**] IS 6937/2, 1983, *Information Processing - Coded character sets for text communication - Part 2: Latin alphabetic and non-alphabetic graphic characters.*

[**13**] IS 7350, 1984, *Text Communication - Registration of graphic character subrepertoires.*

[**14**] IS 8824, 1987, *Information Processing Systems - Open Systems Interconnection - Specification of Abstract Syntax Notation One (ASN.1).*

[**15**] IS 8825, 1987, *Information Processing Systems - Open Systems Interconnection - Specification of Basic Encoding Rules for Abstract Syntax Notation One (ASN.1).*

[**16**] IS 8859-1, 1987, *Information Processing - 8-bit single-byte coded graphic character sets - Part 1: Latin alphabet No. 1.*

[**17**] 1989, *International Register of Coded Character Sets to be Used with Escape Sequences, International Register of Coded Character Sets.*

3 Definitions

3.1 Character data

Character data is defined to be graphic characters and control functions as defined by ISO 2022 and the appropriate International Standards.

3.2 Composite graphic symbol

A composite graphic symbol is defined for the purposes of the International Standardized Profile as a non-spacing diacritical in combination with an alphabetic as in ISO 6937.

4 Abbreviations

4.1 ASN.1

ASN.1 is an abbreviation for Abstract Symbolic Notation One.

4.2 IRV

IRV is an abbreviation for International Reference Version.

5 Position within the taxonomy

The formal position of this International Standardized Profile within the taxonomy is currently unknown.>> It may be referenced from the ISP for any application service element or OSI application.

6 Conformance

Implementations claiming conformance to this ISP must designate one or more of the Character Set Profiles defined herein.

Imaging of Graphic Characters is not required by this ISP. Imaging conformance may be defined in the specific Upper Layers Requirements section of the referencing ISP. If no imaging requirements are specified, then there are no conformance requirements.

6.1 Processed character data

Processed character data is character data which must be processed by the Application Service Element or OSI Application, for example, store and forward character data.

Senders of character data must not produce invalid character codes or invalid designating or invoking escape sequences.

6.1.1 Non-supported character sets

If an implementation receives a designating escape sequence for a character set that it is not able to interpret, then it shall regard that sequence as "invalid data." If possible, it will signal this error in a way that is appropriate to the protocol definition. For applications for which there is no protocol, then no error need be returned. It will not be required to interpret any following characters that are within that data item.

6.1.2 Reserved character codes

If an implementation receives a coded character that is specified in the standard to be "reserved for future standardization," it shall not be considered an error. An imaging device shall indicate receipt of such a reserved character to the user in an implementation defined way, e.g., by making available a character that need not be distinguishable from one of the other characters specified in the standard.

If receivers reject or discard invalid character codes, an appropriate error code must be returned.

6.1.3 Validation of character codes

Character codes within the scope of a standard for which there is no definition in the code table are defined to be invalid character codes. An invalid escape sequence is any designating or invoking escape which is not defined in these agreements.

Implementations must conform to the following statement:

a) Originators of data shall not produce invalid character codes or invalid designating or invoking escape sequences.

6.2 Unprocessed character data

Unprocessed character data is character data which is not processed by the Application Service Element or OSI Application, for example, character matching.

6.2.1 Validation of character codes

Character codes within the scope of a standard for which there is no definition are defined to be invalid character codes. An invalid escape sequence is any designating or invoking escape sequence which is not defined in these agreements.

Implementations must conform to the following statements:

a) Receivers need not validate character codes or designating or invoking escape sequences;

b) Senders who do not originate data need not validate character codes.

7 General agreements

The agreements recorded in this section cover all character set usage except where explicitly noted to the contrary. Additional agreements specific to individual character sets are recorded in the individual character set profiles.

7.1 Encoding

The following agreements cover various aspects of character encoding.

7.1.1 Overprint, composite characters

A composite graphic symbol is considered as one character for purposes of comparison and character string length computation.

With the exception of composite graphic symbols, sequences of graphic characters and control functions which would result in the presentation of two or more graphic characters in a single character position shall not be used. So for example, the sequence "a BACKSPACE" must be processed as three characters rather than as the single character a.

7.1.2 Code extension facilities for GeneralString and GraphicString

This section constitutes the prior agreement on code extension required by ISO 2022.

For ASN.1 GeneralString and GraphicString types, the assumed extension facilities are as though the following escape sequences from ISO 2022 have been applied: ESC 2/0 4/3, ESC 2/0 4/9, and ESC 2/0 5/10. These sequences indicate:

a) 8-bit environment;

b) the G0, and G1 graphic sets shall be used;

c) the designating escape sequences also invoke the G0 and G1 sets into columns 02 and 07 and 10 to 15 respectively;

d) no locking shift functions shall be used;

e) the graphic character sets may comprise 94 and/or 96 characters;

f) a G2 set shall be used;

g) characters from G2 may be accessed by use of the single-shift 2 control function.

Designating ESCAPE sequences in a data stream are permitted. No Announcers of extension facilities may be used within these ASN.1 string types.

7.1.3 Initial conditions for TeletexString

For TeletexString (T61String) the initial condition is described in CCITT T.61 annex A, clause A.2.

7.2 Comparisons

This section contains agreements concerning comparison of characters during processing.

7.2.1 Matching Characters

A character submitted for matching with another character does not have to be drawn from the same coded character set. However, the match is restricted to characters taken from any pair of coded characters sets for which equality or inequality is defined. The identifications of such pairs of coded character sets are shown in the following list. The result of comparing characters from a pair of different coded character sets not in this list is undefined.

a) (ISO 646 ISO 6937-2);

b) (ISO 646, ISO 8859-1);

c) (ISO 6937-2   ISO 8859-1).

Character matching is defined for characters, not their coded representations. The character must take into account any code extension techniques. For example, the character named "SMALL LETTER a WITH DIAERESIS" of ISO 8859 must match the character named "small a with diaeresis or umlaut mark" of ISO 6937 even though the former character is encoded in a single octet and the latter in two octets.

Two characters are said to be equal if, and only if, their names are identical. The names are recorded in the registration of the character sets in the *International Register of Coded Character Sets to be used with Escape Sequences* and not the character set International Standard or Recommendation.

In the case of ISO 6937-2 the names of the composite graphic symbols are specified in the standard itself. However in the present edition there are some systematic differences between the naming conventions used in the standard and those used in the ISO Character Set Register as shown below:

a) ISO 6937 name: capital A with acute accent

b) ISO Register Name: CAPITAL LETTER A WITH ACUTE ACCENT

In this case, two characters are equal if, and only if, their names differ only by the inclusion of the word LETTER in the ISO Register Name. For those characters whose names do not follow this convention, the following list defines the match:

a) ISO 6937 Name ISO Register Name

Editor's Note - <to be filled in>

If a character set registration does not provide character names, then, matching will be defined by exact matching on an octet by octet basis.

Editor's Note - The problem of matching Oriental language character sets is for further study.

In comparing strings, all control functions except code designation and invocation extension facilities shall be ignored. SPACE is treated as a graphic character in such comparisons.

In comparing strings when a character code is encountered for which no other match is defined, matching will be defined by exact matching on an octet by octet basis.

7.2.2 Caseignore comparisons

In character comparisons in which case is ignored, the matching rules of clause 7.2.1 are relaxed in that the characters are equal if their names as defined in clause 7.2.1 differ only by one name having SMALL where the other names has CAPITAL.

7.2.3 Ordering and comparing characters

An agreement on comparison, other than equality or inequality, between characters requires a definition of a collating sequence. This document contains no such agreements.

The collating sequence of letters, accented letters and other graphic symbols is not currently defined in any International Standard or Recommendation.

Preferred collating sequences might vary between countries.

7.2.4 Comparing encoded ASN.1 character strings

In this section, a character string is considered to be a sequence of characters some of which may be composed of multiple bytes depending upon the character set encodings which are specified. Comparing two character strings gives the same result independent of each character string's encoding; for example, the comparison is independent of the Basic Encoding Rules for ASN: 1:

a) as constructed or primitive form, or;

b) as definite or indefinite length form.

8 Character set profiles

A Character Set Profile summarizes implementation agreements specific to a particular character set. Character Set Profiles are identified in the following manner:

CSn-m

where:
CS means Character Set
n = 1 designates a profile for a graphic character set
n = 2 designates a profile for a control function set
m is a number uniquely identifying the Character Set Profile

The values of n and m are defined in this agreement. Names of Character Set Profiles are also defined in this International Standardized Profile.

This section covers agreements about Character Set Standards and Recommendations including:

a) Subrepertoires supported;

b) standardized options selected;

c) component character sets and their registrations in the *International Register of Coded Character Sets to be used with Escape Sequences* where there is a choice to be made, or the standard does not specify it;

d) the designation of component character sets within the ISO 2022 Code Extension Model where there is a choice to be made.

The General Agreements of the preceding section apply to each of these Character Set Profiles.

8.1 CS1-1 ISO 646 Graphic character set

8.1.1 Base standard

International Standard 646 - 1983, *Information Processing - ISO 7-bit coded character set for information interchange*.

Editor's Note - These agreements will be based on the new DIS 646.

8.1.2 Subrepertoire or version

International Reference Version

8.1.3 Standard options selected

Composite graphic symbols are covered by General Agreements.

8.1.4 Character set components and designated position

IRV of ISO 646 number 2 in GO

Editor's Note - This will change to number 6.

Space is in 2/0.

8.1.5 Other agreements

None.

8.2 CS1-2 JIS X0208

Editor's Note - to be defined.

8.3 CS1-3 CCITT Recommendation T.61 graphic character sets basic teletex profiles

8.3.1 Base standard

CCITT Recommendation T.61 - 1985, *Character Repertoire and Coded Character Sets for the International Teletex Service*.

Editor's Note - These references will be updated as soon as the 1989 versions are published.

8.3.2 Subrepertoire or version

None

8.3.3 Standard options selected

None

8.3.4 Character set components and designated position

Teletex Primary Graphic Set 102 in GO.

Teletex Supplementary Graphic Set 103 in G2.

SPACE in 2/0.

8.3.5 Other agreements

Support for CCITT Recommendation T.61 as an ASN.1 GeneralString is outside of this International Standardized Profile.

Support of the graphic set components of T.61 as an ASN.1 GraphicString is outside the scope of this International Standardized Profile.

Use of CCITT Recommendation T.61 except where mandated by standards is outside the scope of this International Standardized Profile. Exceptions to this rule for specific Application Service Element protocol elements must be documented by the referencing Application Service Elements or OSI Applications.

8.4 CS1-4 ISO 8859-1 Latin Alphabet No. 1

8.4.1 Base standard

International Standard 8859-1 - 1987, *Information processing - 8-bit single-byte coded graphic character sets - Part 1 : Latin alphabet No. 1*.

8.4.2 Subrepertoire or version

Not applicable.

8.4.3 Standard options selected

Not applicable.

8.4.4 Character set components and designated position

ASCII Graphic Character Set number 6 in GO.

Right hand part of Latin Alphabet No.1 number 100 in G1.

8.4.5 Other agreements

None.

8.5 CS1-5 ISO 6937-2 Coded character sets for text communication

8.5.1 Base standard

International Standard 6937/2 - 1983, *Information Processing - Coded character sets for text communication - Part 2: Latin alphabetic and non-alphabetic graphic characters*.

> **Editor's Note -** Includes Addendum 1 as soon as it is published.

8.5.2 Subrepertoire or version

Full number 0.

Minimum number 1.

Telenex number 3.

Western European Data Processing number 9.

8.5.3 Standard options selected

Not applicable

8.5.4 Character set components and designated position

IRV of ISO 646 number 2 in GO.

Editor's Note - This will change to number 6.

Supplementary set of Latin Text Processing number 142 in G2.

8.5.5 Other agreements

For subrepertoires 2 and 5, the supplementary set may be omitted at the discretion of the sending application.

8.6 CS1-6 ISO 8859/7 Greek supplementary set

Editor's Note - to be defined.

8.7 CS1-7 CCITT Recommendation T.61 Graphic character sets basic teletex profiles (1984)

8.7.1 Base standard

CCITT Recommendation T.61 - 1981, *Character Repertoire and Coded Character Sets for the International Teletex Service*.

8.7.2 Subrepertoire or version

None

8.7.3 Standard options selected

None

8.7.4 Character Set Components and Designated Position

Teletex Primary Graphic Set 102 in GO.

Teletex Supplementary Graphic Set 103 in G2.

SPACE in 2/0.

8.7.5 Other agreements

Support for CCITT Recommendation T.61 as an ASN.1 GeneralString is outside of this International Standardized Profile.

Support of the graphic set components of T.61 as an ASN.1 GrahpicString is outside the scope of this International Standardized Profile.

Use of CCITT Recommendation T.61 except where mandated by standards is outside the scope of this International Standardized Profile. Exceptions to this rule for specific Application Service Element protocol elements must be documented in the referencing Application Service Elements or OSI Applications.

This profile is intended for use with the X.400-1984 Implementation agreements only.

8.8 CS 1-8 CCITT Recommendation T.61 Graphic character sets

Editor's Note - to be defined.

8.9 Korean National character set

Editor's Note - to be defined

8.10 CS2-1 ISO 646 CO control functions

8.10.1 Base standard

International Standard 646-1983, *Information Processing - ISO 7-bit coded character set for information interchange*.

8.10.2 Subrepertoire or version

None

8.10.3 Standard options selected

None

8.10.4 Character set components and designated position

ISO 646 CO Set number 1 in CO.

DELETE in 7/15.

8.10.5

When a single format effector for vertical (or horizontal) movement is optionally permitted to effect a combined vertical and horizontal movement, implementations shall not use this single format effector for effecting the combined vertical and horizontal movement.

8.11 CS2-2 ISO 6429 additional control functions

8.11.1 Base standard

International Standard 6429 - 1983, *Information Processing - ISO 7-bit and 8-bit coded character sets - Additional control functions for character-imaging devices*.

8.11.2 Subrepertorie or version

None

8.11.3 Standard options selected

None

8.11.4 Character set components and designated position

C1 Control Set of ISO 6429 - 1983 number 77 in C1.

8.11.5 Other agreements

None

8.12 CS2-3 CCITT Recommendation T.61 Control Sets

8.12.1 Base standard

CCITT Recommendation T.61 - 1985, *Character Repertoire and Coded Character Sets for the International Teletex Service*.

> **Editor's Note -** These references will be updated as soon as the 1989 versions are published.

8.12.2 Subrepertoire or version

None

8.12.3 Standard options selected

Teletex optional repertoire of control functions is not selected.

8.12.4 Character set components and designated position

Teletex Primary Set of Control Functions number 106 in C0.

Teletex Supplementary Set of Control Functions number 107 in C1.

8.12.5 Other agreements

None

8.13 CS2-4 CCITT Recommendation T.61 Control sets (1984)

8.13.1 Base standard

CCITT Recommendation T.61 - 1981, *Character Repertoire and Coded Character Sets for the International Teletex Service*.

8.13.2 Subrepertoire or version

None

8.13.3 Standard options selected

Teletex optional repertoire of control functions is not selected.

8.13.4 Character set components and designated position

Teletex Primary set of Control Functions number 106 in C0.

Teletex Supplementary Set of Control Functions number 107 in C1.

8.13.5 Other agreements

This profile is intended for use with the X.400-1984 implementation agreements only.

Annex A (informative)

Character Set Technology

(This annex does not form part of these agreements.)

A.1 Introduction

This Annex presents information from Information Processing Character Set Standards which is relevant to the implementation of OSI Services. The intent is to collect into one place the most relevant information for implementors from character set standards specified in OSI and OSI related standards.

A.2 Scope

Material in this Annex is drawn from ISO and CCITT Character Set standards and Recommendations. Topics covered include character Set Extension Techniques and Character Set Encodings. ASN.1 Basic Encoding Rules are reviewed also. Rationale for the implementation agreements in the ISP is provided where appropriate.

A.3 Field of application

This annex covers character set information for ASN.1 Basic Encoding Rules are used by OSI services. It also includes information pertaining to OSI Interchange Formats such as Office Document Architecture.

A.4 Character set standards

The following character set standards have some relevance to this material.

International Information Exchange for Videotex, CCITT Recommendation T.100, 1985.

International Alphabet No. 5, CCITT Recommendation T.50, 1985.

Coded Character Sets for Telematic Services, CCITT Recommendation T.51, 1985.

Character Repertoire and Coded Character Sets for the International Teletex Service, CCITT Recommendation T.61, 1985.

Information processing - 8-bit single-byte coded graphic character sets - Part 7: Latin/Greek Alphabet, DIS 8859-7, 1987.

Information Processing - ISO 7-bit and 8-bit coded character - Code extension techniques, IS 2022, 1986.

Data processing - Procedure for registration of escape sequences, IS 2375, 1985.

Information processing - ISO 8-bit code for information interchange - Structure and rules for implementation, IS 4873, 1986

Information Processing - ISO 7-bit and 8-bit coded character sets - Additional control functions for character-imaging devices, IS 6429, 1983.

Information Processing - ISO 7-bit coded character set for information interchange, IS 646, 1983.

Information Processing - Coded character sets for text communication - Part 1: General introduction, IS 6937/1, 1983.

Information Processing - Coded character sets for text communication - Part 2: Latin alphabetic and non-alphabetic graphic characters, IS 6937/2, 1983.

Text communication -Registration of graphic character subrepertoires, IS 7350, 1984.

Information Processing Systems - Open Systems Interconnection - Specification of Abstract Syntax Notation One (ASN.1), IS 8824, 1987.

Information Processing Systems - Open Systems Interconnection - Specification of Basic Encoding Rules for Abstract Syntax Notation One (ASN.1), IS 8825, 1987.

Information processing - 8 bit single-byte coded graphic character sets - Part 1:Latin Alphabet No. 1, IS 8859-1, 1987.

International Register of Coded Character Sets to be used with Escape Sequences, International Register of Coded Character Sets, 1989.

A.5 Introduction to character set standards

A brief introduction to reading a character set standard is presented here for the uninitiated. Most of the character set standards described in this Annex use the term "bit combinations" to refer to the ordered string of bits which compose a character. Most implementations of these standards allocate an 8-bit byte to a character and consequently tend to intermix the notions of bytes and characters. In the OSI environment, 8-bit combinations are normally referred to as "octets."

A character set standard generally presents its character encodings in a table composed of 16 rows and 8 or 16 columns depending on whether a 7-bit or an 8-bit character set is being defined. A given character code is generally referenced by naming its column and then its row. Thus in ISO 646 the capital letter A is referred to as 4/1. Some standards precede single digits with a zero so that in ISO 8859/1 the capital letter A is referred to as 04/01. This positional notation is especially important in the consideration of the code extension techniques. Code extension techniques describe characters in terms of their position only, without regard for any possible previously assigned interpretations.

A.6 Definitions

The following definitions drawn from relevant character set standards are provided to assist in understanding the material in this annex. These definitions were drawn from International Standards which were current at the time of drafting this document. Any conflict between these definitions and those of the relevant international standards shall be resolved by using the definition in the International Standard.

a) bit combination: An ordered set of bits that represents a character or is used as a part of the representation of a character.

b) byte: A bit string that is operated upon as a unit and the size of which is independent of redundancy or framing techniques.

c) character: A member of a set of elements used for the organization, control or representation of data.

d) code extension: The techniques for the encoding of characters that are not included in the character set of a given code.

e) control character: A control function where the coded representation of which consists of a single bit combination.

f) control function: An action that affects the recording, processing, transmission or interpretation of data and that has a coded representation consisting of one or more bit combinations.

g) graphic character: A character, other than a control function, that has a visual representation normally handwritten, printed or displayed.

A.7 ISO 2022 Information Processing - iso 7-bIT AND 8-bIT coded character sets - code extension techniques

This International Standard was originally written to establish extension techniques for the 7-bit codes of ISO 646. It has been revised twice so that is now also provides the basic framework for an 8-bit code family which is compatible with the 7-bit codes. The four interrelated clauses cover:

a) the extension of the 7-bit code remaining in a 7--bit environment;

b) the structure of a family of 8-bit codes;

c) the extension of an 8-bit code remaining in an 8-bit environment;

d) the relationship between the 7-bit code and an 8-bit code.

The middle two clauses are of special relevance to this document although portions of the others should be read and understood in order to set the context for the relevant material.

Some underlying assumptions from the standard are recorded here in order to understand the context of these agreements. Clause 2 notes that code extension techniques are designed to be used for data to be processed serially in a forward direction.

A.7.1 Structure of a family of 8-bit codes

Clause 7 of the standard describes a family of 8-bit codes obtained from the 7-bit set. The family of 8-bit codes is obtained by the addition of one bit to each of the bit combinations of the 7-bit code producing a set of 256 8-bit combinations. The characters of the 7-bit code are assigned to the 128 bit combinations for which the eighth bit is set to ZERO. The 128 additional bit combinations for which the eighth bit is set to ONE are available for assignment. The 8-bit code table of clause 7.1 is a 16 by 16 array of columns numbered 00 to 15 and rows numbered 0 to 15. Columns 08 and 09 are provided for control characters and columns 10 to 15 for graphic characters.

The following figure shows the basic code structure for 8-bit character codes. This structure is followed by the standards described in this annex.

<table>
<tr><th></th><th>00</th><th>01</th><th>02</th><th>03</th><th>04</th><th>05</th><th>06</th><th>07</th><th>08</th><th>09</th><th>10</th><th>11</th><th>12</th><th>13</th><th>14</th><th>15</th></tr>
<tr><td>0</td><td colspan="2" rowspan="16">A set of 32 control characters</td><td>SP</td><td colspan="5"></td><td colspan="2" rowspan="16">A set of 32 control characters</td><td>10/0</td><td colspan="5"></td></tr>
<tr><td>1</td><td colspan="6" rowspan="14">A set of 94 or 96 graphic characters</td><td colspan="6" rowspan="14">A set of 94 or 96 graphic characters</td></tr>
<tr><td>2</td></tr>
<tr><td>3</td></tr>
<tr><td>4</td></tr>
<tr><td>5</td></tr>
<tr><td>6</td></tr>
<tr><td>7</td></tr>
<tr><td>8</td></tr>
<tr><td>9</td></tr>
<tr><td>10</td></tr>
<tr><td>11</td></tr>
<tr><td>12</td></tr>
<tr><td>13</td></tr>
<tr><td>14</td></tr>
<tr><td>15</td><td colspan="5"></td><td>DEL</td><td colspan="5"></td><td>15/15</td></tr>
</table>

Figure A1 - 8-bit Code Structure

The family concept is described in clause 7.2 as follows:

a) a set of 32 additional control characters can be selected for columns 08 and 09;

b) a set of 94 or 96 additional graphic characters can be selected for columns 10 to 15. If a set of 94 graphic characters is invoked in columns 10 to 15, positions 10/0 and 15/15 shall not be used.

Three control functions were provided by ISO 646 for purposes of code extension. ISO 2022 uses these three and adds 7 more for use in the 8-bit environment. For reference purposes, the corresponding characters from the 7-bit environment are shown also. The following table shows these control functions.

Table A1 - Control functions

7-bit Name	**Abbreviation**	**8-bit Name**	**Abbreviation**
ESCAPE	ESC	ESCAPE	ESC
SHIFT-OUT	SO	LOCKING-SHIFT ZERO	LS0
SHIFT-IN	S1	LOCKING-SHIFT ONE	LS1
LOCKING-SHIFT TWO	LS2	LOCKING-SHIFT TWO	LS2
LOCKING-SHIFT THREE	LS3	LOCKING-SHIFT THREE	LS3
SINGLE-SHIFT TWO	SS2	SINGLE-SHIFT TWO	SS2
SINGLE-SHIFT THREE	SS3	SINGLE-SHIFT THREE	SS3
		LOCKING-SHIFT ONE RIGHT	LS1R
		LOCKING-SHIFT TWO RIGHT	LS2R
		LOCKING-SHIFT THREE RIGHT	LS3R

A.7.2 Elements of code extension in an 8-bit environment

The elements of code extension in an 8-bit environment are shown in the following table taken from clause 8.1 of the standard.

Table A2 - Elements of code extension in an 8-bit environment

Set	Description	Columns occupied
C0	32 control characters	00 to 01
C1	32 control characters	08 to 09
G0	94 graphic characters	02 to 07
G1	94 or 96 graphic characters	02 to 07 or 10 to 15
G2	94 or 96 graphic characters	02 to 07 or 10 to 15
G3	94 or 96 graphic characters	02 to 07 or 10 to 15

A.7.3 Multiple character sets

Editor's Note - Describe multi-level designation and invocation here.

The standard defines a graphic character set extension strategy in which a designating escape sequence is used to select up to four graphic character sets from the International Character Set Register. An invocation sequence is then used to select up to two graphic sets from the designated sets for concise assess to the characters. The following figure shows the technique for the 8-bit environment.

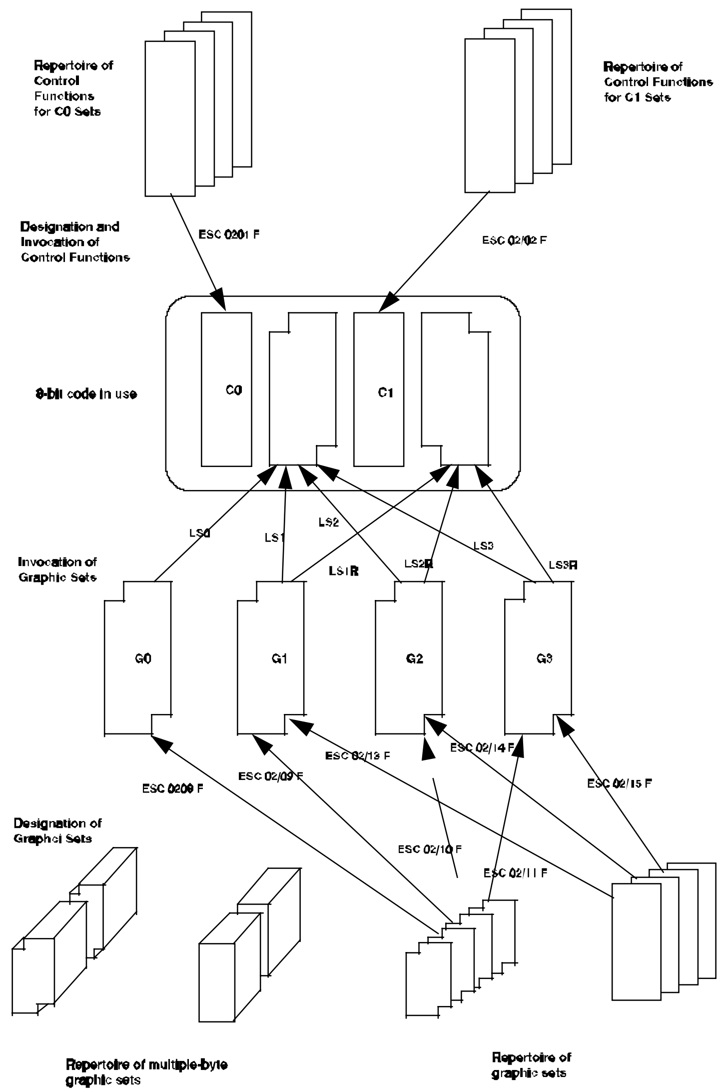

Figure A2 - Code Extension in an 8-bit Environment

The standard defines two terms for use in describing code extension practices: to designate and to invoke. They are defined as follows:

a) to designate: To identify a set of characters that are to be represented, in some cases immediately and in others on the occurrence of a further control function, in prescribed manner.

b) to invoke: To cause a designated set of characters to be represented by the prescribed bit combinations whenever those bit combinations occur, until an appropriate code extension function occurs.

Designation of a character set is usually achieved by employing an escape sequence defined by the standard along with values assigned by a registration authority. In many cases, designation of a character set also implies invocation. In other cases, a character set must be explicitly invoked usually by using a shift function.

The following table defines the use of the locking shift functions in an 8-bit environment for extension of the graphic set.

Table A3 - Locking shift functions in an 8-bit environment for extension

Function	Abbreviation	Set invoked	Columns affected
LOCKING-SHIFT ZERO	LS0	G0	02 to 07
LOCKING-SHIFT ONE	LS1	G1	02 to 07
LOCKING-SHIFT ONE RIGHT	LS1R	G1	10 to 15
LOCKING-SHIFT TWO	LS2	G2	02 to 07
SINGLE-SHIFT TWO RIGHT	LS2R	G2	10 to 15
LOCKING-SHIFT THREE	LS3	G3	02 to 07
LOCKING-SHIFT THREE RIGHT	LS3R	G3	10 to 15

The meanings of control characters in columns 00, 01, 08 and 09 will not be affected by the occurrence of these locking shift functions.

Clause 6.4 states that at the beginning of any information interchange, except where interchanging parties have agreed otherwise, all designations shall be defined by the use of appropriate escape sequences, and the shift status shall be defined by the use of the appropriate locking shift functions.

A.7.4 Announcement of extension facilities

A code extension facility consists of the elements of code extension employed as well as the means by which these elements are designated and invoked. Thus the control function sets, the graphic character sets, and the character shifting codes must be specified. Specification of control function sets and graphic character sets also specifies the designation and invocation sequences required to use their codes.

Clause 9 of ISO 2022 describes how the various extension facilities are to be made known. If an announcement is to be embedded in the interchanged information, the form is described. The announcement may be omitted by agreement between the interchanging parties. Some restrictions are imposed on the defined announcer sequences. For example the sequence ESC 02/00 04/03 specifies that 1) the G0 and G1 sets shall be used in an 8-bit environment only; 2) the designating escape sequences also invoke the G0 and G1 sets into columns 02 to 07 and 10 to 15; respectively, and 3) no locking shift functions shall be used.

A.7.5 Composite graphic characters

Clause 6.1.8 of the standard addresses methods for the representation of additional graphic characters by the combination of two or more graphic characters in the same position. Two methods are provided for:

a) graphic characters having implicit forward motion (spacing characters) used in conjunction with BACKSPACE or CARRIAGE RETURN;

b) graphic characters having no implicit forward motion (non-spacing characters) used in combination with spacing graphic characters.

Method b allows for the specification of characters with diacritical marks. The technique is known colloquially as the "dead key" approach. An non-spacing accent grave character is immediately followed by the character it modifies.

A.7.6 International register of coded character sets to be used with escape sequences

ISO 2375 specifies procedures to be used to assign meanings to the final bit combinations of escape sequences defined in ISO 2022. The International Register of Coded Character Sets to be used with escape sequences is the document which records these assignments. The current International Registration Authority for ISO 2375 is the European Computer Manufacturers Association (ECMA).

A.8 Character sets

Several character set standards are described here. The standards chosen for description are each used by one or more OSI applications. The usage of these standards is summarized in tabular form.

A.8.1 ISO 646 7-bit coded character set for information processing interchange and CCITT Recommendation T.50 international alphabet no. 5

This International Standard specifies a set of 128 characters with their coded representation. The 128 bit combinations of the 7-bit code represent control characters and graphic characters. The allocation of characters to bit combinations is based on the following principles:

a) the bit combinations 0/0 to 1/15 represent 32 control characters;

b) the bit combination 2/0 represents the character SPACE, which is interpreted as both a control character and a graphic character;

c) the bit combinations 2/1 to 7/14 represent up to 94 graphic characters;

d) the bit combination 7/15 represents the control character DELETE.

The 7-bit code table consists of 128 positions arranged in 8 columns and 16 rows. The columns are numbered from 0 to 7, and rows are numbered 0 to 15.

Most of these characters are mandatory and unchangeable, but provision is made for some flexibility to accommodate national and other requirements. The standard provides guidance on how to exercise the options offered in order to define specific national versions and application-oriented versions. It further specifies an International Reference Version in which all options have been exercised.

Editor's Note - A revision of ISO 646 which has achieved DP status revises this table.

Table A4 - X3.4 - 1977 ASCII

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**
0	NUL	DEL	SP	0	@	P	‘	p
1	SOH	DC1	!	1	A	Q	a	q
2	SIX	DC2	"	2	B	R	b	r
3	ETX	003	#	3	C	S	c	s
4	EDT	DC4	$	4	D	T	d	t
5	**ENQ**	NAK	%	5	E	U	e	u
6	ACK	SYN	&	6	F	V	f	v
7	BEL	ET6	’	7	G	W	g	w
8	BS	CAN	(	8	H	X	h	x
9	HT	EM	)	9	I	Y	i	y
10	LF	SUB	*	:	J	Z	j	z
11	VT	ESC	+	;	K	[	k	{
12	FF	FS	,	<	L	\	l	|
13	CR	GS	.-	=	M	]	m	}
14	S0	RS	.	>	N	^	n	"
15	S1	US	/	?	O	__	o	DEL

Table A5 - ISO 646 - 1983 IRV

	0	1	2	3	4	5	6	7
0	NUL	TC7	SP	0	@	P	‘	p
1	TC1	DC1	!	1	A	Q	a	q
2	RC2	DC2	"	2	B	R	b	r
3	TC3	DC3	#	3	C	S	c	s
4	TC4	DC4	$	4	D	T	d	t
5	**TC5**	TC8	%	5	E	U	e	u
6	TC6	TC9	&	6	F	V	f	v
7	BEL	TC10	’	7	G	W	g	w
8	FED	CAN	(	8	H	X	h	x
9	FE1	EM	)	9	I	Y	i	y
10	FE2	SUB	*	:	J	Z	j	z
11	FE3	ESC	+	;	K	[	k	{
12	FE4	IS4	,	<	L	\	l	|
13	FE5	IS3	.-	=	M	]	m	}
14	S0	IS2	.	>	N	^	n	"
15	S1	IS1	/	?	O	__	o	DEL

ISO 646 International Reference

A.8.2 ISO 8859 Information processing - 8-bit single-byte coded character sets

This International Standard is a multiple part standard. Each part specifies a set of up to 191 graphic characters and the coded representation of each of these characters by means of a single 8-bit byte. The use of control functions for the coded representation of composite characters is prohibited. Each set is intended for a group of languages. Part 1 of ISO 8859 specifies a set of 191 graphic characters identified as Latin alphabet No. 1. This set of graphic characters is suitable for use in a version of an 8-bit code according to ISO 2022.

The standard specifically notes that it is not intended for use with CCITT defined Telematic services. If information coded according to ISO 8859 is to be transferred to such services, it will have to conform at the coding interface to their requirements.

Table A6 - ISO 8859/1- 1987 Latin Alphabet No. 1

<table>
<tr><th></th><th>0</th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th><th>7</th><th>8</th><th>9</th><th>10</th><th>11</th><th>12</th><th>13</th><th>14</th><th>15</th></tr>
<tr><td>0</td><td></td><td></td><td>SP</td><td>0</td><td>@</td><td>P</td><td>‘</td><td>p</td><td></td><td></td><td>MBSP</td><td>º</td><td>À</td><td>Đ</td><td>à</td><td>đ</td></tr>
<tr><td>1</td><td></td><td></td><td>!</td><td>1</td><td>A</td><td>Q</td><td>a</td><td>q</td><td></td><td></td><td>¡</td><td>±</td><td>Á</td><td>Ñ</td><td>á</td><td>ñ</td></tr>
<tr><td>2</td><td></td><td></td><td>"</td><td>2</td><td>B</td><td>R</td><td>b</td><td>r</td><td></td><td></td><td>¢</td><td>²</td><td>Â</td><td>Ò</td><td>â</td><td>ò</td></tr>
<tr><td>3</td><td></td><td></td><td>#</td><td>3</td><td>C</td><td>S</td><td>c</td><td>s</td><td></td><td></td><td>£</td><td>³</td><td>Ã</td><td>Ó</td><td>ã</td><td>ó</td></tr>
<tr><td>4</td><td></td><td></td><td>$</td><td>4</td><td>D</td><td>T</td><td>d</td><td>t</td><td></td><td></td><td>¤</td><td>'</td><td>Ä</td><td>Ô</td><td>ä</td><td>ô</td></tr>
<tr><td>5</td><td></td><td></td><td>%</td><td>5</td><td>E</td><td>U</td><td>e</td><td>u</td><td></td><td></td><td>¥</td><td>µ</td><td>Ă</td><td>Õ</td><td>ă</td><td>õ</td></tr>
<tr><td>6</td><td></td><td></td><td>&</td><td>6</td><td>F</td><td>V</td><td>f</td><td>v</td><td></td><td></td><td>|</td><td>π</td><td>Æ</td><td>Ö</td><td>æ</td><td>ö</td></tr>
<tr><td>7</td><td></td><td></td><td>’</td><td>7</td><td>G</td><td>W</td><td>g</td><td>w</td><td></td><td></td><td>§</td><td>•</td><td>Ç</td><td>×</td><td>ç</td><td>÷</td></tr>
<tr><td>8</td><td></td><td></td><td>(</td><td>8</td><td>H</td><td>X</td><td>h</td><td>x</td><td></td><td></td><td>¨</td><td>,</td><td>È</td><td>Ø</td><td>è</td><td>ø</td></tr>
<tr><td>9</td><td></td><td></td><td>)</td><td>9</td><td>I</td><td>Y</td><td>i</td><td>y</td><td></td><td></td><td>©</td><td>¹</td><td>É</td><td>Ù</td><td>é</td><td>ù</td></tr>
<tr><td>10</td><td></td><td></td><td>*</td><td>:</td><td>J</td><td>Z</td><td>j</td><td>z</td><td></td><td></td><td>[illegible]</td><td>O</td><td>Ê</td><td>Ú</td><td>ê</td><td>ú</td></tr>
<tr><td>11</td><td></td><td></td><td>+</td><td>;</td><td>K</td><td>[</td><td>k</td><td>{</td><td></td><td></td><td>«</td><td>»</td><td>Ë</td><td>Û</td><td>ë</td><td>û</td></tr>
<tr><td>12</td><td></td><td></td><td>,</td><td><</td><td>L</td><td>\</td><td>l</td><td>|</td><td></td><td></td><td>¬</td><td>¼</td><td>Ì</td><td>Ü</td><td>ì</td><td>ü</td></tr>
<tr><td>13</td><td></td><td></td><td>.-</td><td>=</td><td>M</td><td>]</td><td>m</td><td>}</td><td></td><td></td><td>SHY</td><td>½</td><td>Í</td><td>Ý</td><td>í</td><td>ý</td></tr>
<tr><td>14</td><td></td><td></td><td>.</td><td>></td><td>N</td><td>^</td><td>n</td><td>"</td><td></td><td></td><td>®</td><td>¾</td><td>Î</td><td>Þ</td><td>î</td><td>þ</td></tr>
<tr><td>15</td><td></td><td></td><td>/</td><td>?</td><td>O</td><td>__</td><td>o</td><td>DEL</td><td></td><td></td><td>—</td><td>¿</td><td>Ï</td><td>ß</td><td>ï</td><td>ÿ</td></tr>
</table>

ISO 8859/1 - 1987 Latin Alphabet No. 1

A.8.3 ISO 6937 Information processing - coded character sets for text communication

This International Standard specifies repertoires of graphic characters and control functions, and their coded representation for use in text communication. This International Standard consists, at present, of two parts as follows:

a) ISO 6937/1, General Introduction;

b) ISO 6937/2, Latin Alphabetic and non-alphabetic graphic characters.

The specifications are based on the 7-bit coded character set specified in ISO 646, the 7-bit and 8-bit code extension techniques of ISO 2022, and the definitions of additional control functions given in ISO

6429.

ISO 6937 was developed in parallel with CCITT Recommendations which in the standard are referred to as S.61 and S.100. These CCITT Recommendations were moved to a new section in 1984 and were renumbered T.61 and T.100. This 1984 designation is being carried forward in the 1988 CCITT Recommendations.

A.8.3.1 ISO 6937/1 Information processing - coded character sets for text communication - Part 1: general introduction

Annex A of this international Standard describes a method of identification of graphic characters and control functions which is used in other parts of the standard to define the characters of the standard.

A.8.3.2 ISO 6937/2 Information processing - coded character sets for text communication - Part 2: Latin alphabetic and non-alphabetic graphic characters

This part of the standard:

a) defines a repertoire of Latin alphabetic and non-alphabetic characters for the communication of text in European languages;

b) specifies coded representations for the graphic characters;

c) specifies rules for the definition and use of graphic character subrepertoires.

A graphic subrepertoire is a subset of the defined character repertoire. Because the number of characters defined by this standard is so large, this subsetting facility allows for the use of well defined subsets of the characters available. Rules for the definition of subrepertoires are defined in clause 5. The procedure for registration of subrepertoires is given in ISO 7350. Three standard subrepertoires are defined in Annex A of the standard.

Graphic characters which represent accented letters and umlauts are specified using a two byte sequence composed of the diacritical character immediately followed by the character modified. The allowable combinations are carefully defined in the standard and only these combinations are permitted.

Table A7 - ISO 6937/2 1983 addendum 1
full repertoire

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0			SP	0	@	P	‘	p			MBSP	○		—	Ω	κ
1			!	1	A	Q	a	q			¡	±	`	¹	Æ	æ
2			"	2	B	R	b	r			¢	²	´	®	Đ	đ
3			#	3	C	S	c	s			£	³	ˆ	©	Õ	õ
4			¤	4	D	T	d	t			$	×	˜	™	Ħ	ħ
5			%	5	E	U	e	u			¥	µ	¯	♪		’
6			&	6	F	V	f	v				π	˘	¬	Ū	ű
7			’	7	G	W	g	w			§	•	˚	|	Ŀ	ŀ
8			(	8	H	X	h	x				÷	¨		Ł	ł
9			)	9	I	Y	i	y			‘	'			Ø	ø
10			*	:	J	Z	j	z			“	”	º		Œ	œ
11			+	;	K	[	k	{			«	»	≺			ß
12			,	<	L	\	l	|			←	¼	_	⅛	Þ	þ
13			.-	=	M	]	m	}			↑	½	˝	⅜	Ŧ	ŧ
14			.	>	N	^	n	"			→	¾	˛	⅝	Ŋ	ŋ
15			/	?	O	_	o	DEL			↓	¿	ˇ	⅞	ń	SHY

A.8.4 CCITT Recommendation T.51 Coded Character Sets for Telematic Services

This recommendation specifies a primary set and a supplementary set of graphic characters which are to be the respective supersets of various primary and supplementary character sets to be used in various telematic services. The Recommendation also describes those code extension mechanism which are relevant to existing telematic services.

A.8.5 CCITT Recommendation T.61 Character Repertoire and Coded Character Sets for the International Teletex Service

This Recommendation contains detailed definitions of the repertoires of graphic characters and control functions to be used in the basic International Teletex service, and their coded representations for communication.

A.9 ASN.1 character String Types

Character String Types are sequences of zero, one or more characters from some specified character set. ISO 8824 defines 8 such types: NumericString, PrintableString, TeletexString (T61String), VideotexString, VisibleString (ISO646String), 1A5String, GraphicString, GeneralString.

A.9.1 Universal Class Numbers and Registration Numbers

The type of each character string is identified by a Universal Class number. Universal Class numbers are assigned by ISO 8824. No other standard or private user may define these numbers. The character sets associated with each type are identified by the ISO Character Set Registration Numbers as shown in the following table:

Table A8 - ISO Character Set Registration Numbers

Name of Character String Type	Universal Class Number	ISO Character Set Registration Number
NumericString	18	Not Registered
PrintableString	19	Not Registered
TeletexString (T61String)	20	87, 102, 103, 106, 107 + SPACE + DELETE
Visible String (ISO646String)	21	1, 72, 73, 102, 108, 128, 129 + SPACE + DELETE
IA5String	26	2 + SPACE
GraphicString	22	1, 2 + SPACE + DELETE
GeneralString	25	All G sets + SPACE
	27	All G sets and all C sets + SPACE + DELETE

NumericString and PrintableString do not have registration Numbers assigned to them since their character sets are defined in table 4 and 5 respectively of ISO 8824.

A.9.2 Initial States

Some character string types allow multiple character sets through code extension techniques. For these types, at the beginning of each string there are initial default character sets to be designated in G0 and/or C0 and/or C1 and for each character set there is an assumed escape sequence. The following table drawn from ISO 8825 describes these initial states.

Table A9 - Initial States

Name of Character String Type	Initial G0(Reg. No.)	Initial C0 (Reg. No.)	Initial C1 (Reg. No.)	Initial ESC Seq and Lock Shift Function	Code Extension
NumericString	2	None	None	ESC 2/8 4/0 LSO	No
PrintableString	2	None	None	ESC 2/8 4/0 LSO	No
TeletexString (T61String)	102	106	107	ESC 2/8 4/0 LSO ESC 2/1 4/5 ESC 2/2 4/8	Yes
VideotexString	102	1	73	ESC 2/8 7/5 LS0 ESC 2/1 4/0 ESC 2/2 4/1	Yes
VisibleString (ISO646String)	2	None	None	ESC 2/8 4/0 LS0	No
IA5String	2	1	None	ESC 2/8 4/0 LS0 ESC 1/1 4/0	No
GraphicString	2	None	None	ESC 2/8 4/0 LS0	Yes
GeneralString	2	1	None	ESC 2/1 4/0 LS0 ESC 2/1 4/0	Yes

For example, VideotexString initial G0 set is Primary Teletex Graphic Set (ISO Registration Number 102), initial C0 set is ISO 646 C0 set (ISO Registration Number 1), initial C1 set is attribute Control Set for Videotex (ISO Registration number 73), initial escape sequence and locking shift function is ESC 2/8/7/5 LSO, and ESC 2/2 4/1, and code extensions are permitted.

A.10 Use of ASN.1 OcteString as a Character String

Editor's Note - Add a description of ODA treatment of character sets

A.11 Escape Sequences for Character Set Designation

This information is extracted from the ISO Register. In some cases, the defaults supplied by ASN.1 make the use of these escape sequences unnecessary. In some cases, this information is carried by application protocol elements.

Table A10 - Graphic Set Designation

Set No.	G0	G1	G2	Name
2	ESC 2/8 4/0			ISO 646 IRV
6	ESC 2/8 4/2			ISO 646 USA
87	ESC 2/4 2/8 4/2	ESC 2/4 2/9 4/2		JIS X0208
100		ESC 2/13 4/1	ESC 2/14 4/1	ISO 8859/1 Right Hand Part
102	ESC 2/8 7/5			CCITT T.61 Primary
103				CCITT 5.61 Supp
126		ESC 2/13 4/6		ISO 8859/7 Greek
142			ESC 2/14 4/10	ISO 6937/2 Ad1 Supp

Table A11 - Control Set Designation

Set No.	C0	C1	Name
1	ESC 2/1 4/0		ISO 646 C0
106	ESC 2/1 4/5		CCITT 5.61 Primary
107		ESC 2/2 4/8	CCITT 5.61 Suppl.

Editor's Note - Add 6429 designation.

Editor's Note - Add DIS 10538 and DIS 10367?