U.S. Department of Commerce

National Bureau of Standards

Computer Science and Technology

NBS Special Publication 500-81

A Survey of Standardization Efforts of Coded Character Sets for Text Processing

NATIONAL BUREAU OF STANDARDS

The National Bureau of Standards[1] was established by an act of Congress on March 3, 1901. The Bureau's overall goal is to strengthen and advance the Nation's science and technology and facilitate their effective application for public benefit. To this end, the Bureau conducts research and provides: (1) a basis for the Nation's physical measurement system, (2) scientific and technological services for industry and government, (3) a technical basis for equity in trade, and (4) technical services to promote public safety. The Bureau's technical work is performed by the National Measurement Laboratory, the National Engineering Laboratory, and the Institute for Computer Sciences and Technology.

THE NATIONAL MEASUREMENT LABORATORY provides the national system of physical and chemical and materials measurement; coordinates the system with measurement systems of other nations and furnishes essential services leading to accurate and uniform physical and chemical measurement throughout the Nation's scientific community, industry, and commerce; conducts materials research leading to improved methods of measurement, standards, and data on the properties of materials needed by industry, commerce, educational institutions, and Government; provides advisory and research services to other Government agencies; develops, produces, and distributes Standard Reference Materials; and provides calibration services. The Laboratory consists of the following centers:

Absolute Physical Quantities[2] — Radiation Research — Thermodynamics and Molecular Science — Analytical Chemistry — Materials Science.

THE NATIONAL ENGINEERING LABORATORY provides technology and technical services to the public and private sectors to address national needs and to solve national problems; conducts research in engineering and applied science in support of these efforts; builds and maintains competence in the necessary disciplines required to carry out this research and technical service; develops engineering data and measurement capabilities; provides engineering measurement traceability services; develops test methods and proposes engineering standards and code changes; develops and proposes new engineering practices; and develops and improves mechanisms to transfer results of its research to the ultimate user. The Laboratory consists of the following centers:

Applied Mathematics — Electronics and Electrical Engineering[2] — Mechanical Engineering and Process Technology[2] — Building Technology — Fire Research — Consumer Product Technology — Field Methods.

THE INSTITUTE FOR COMPUTER SCIENCES AND TECHNOLOGY conducts research and provides scientific and technical services to aid Federal agencies in the selection, acquisition, application, and use of computer technology to improve effectiveness and economy in Government operations in accordance with Public Law 89-306 (40 U.S.C. 759), relevant Executive Orders, and other directives; carries out this mission by managing the Federal Information Processing Standards Program, developing Federal ADP standards guidelines, and managing Federal participation in ADP voluntary standardization activities; provides scientific and technological advisory services and assistance to Federal agencies; and provides the technical foundation for computer-related policies of the Federal Government. The Institute consists of the following centers:

Programming Science and Technology — Computer Systems Engineering.

[1] Headquarters and Laboratories at Gaithersburg, MD, unless otherwise noted;
mailing address Washington, DC 20234.

[2] Some divisions within the center are located at Boulder, CO 80303.

Computer Science and Technology

NBS Special Publication 500-81

A Survey of Standardization Efforts of Coded Character Sets for Text Processing

Joan E. Knoerdel

Center for Computer Systems Engineering
Institute for Computer Science and Technology
National Bureau of Standards
Washington, DC 20234

U.S. DEPARTMENT OF COMMERCE
Malcolm Baldrige, Secretary

National Bureau of Standards
Ernest Ambler, Director

Issued September 1981

Reports on Computer Science and Technology

The National Bureau of Standards has a special responsibility within the Federal Government for computer science and technology activities. The programs of the NBS Institute for Computer Sciences and Technology are designed to provide ADP standards, guidelines, and technical advisory services to improve the effectiveness of computer utilization in the Federal sector, and to perform appropriate research and development efforts as foundation for such activities and programs. This publication series will report these NBS efforts to the Federal computer community as well as to interested specialists in the academic and private sectors. Those wishing to receive notices of publications in this series should complete and return the form at the end of this publication.

National Bureau of Standards Special Publication 500-81
Nat. Bur. Stand. (U.S.), Spec. Publ. 500-81, 50 pages (Sept. 1981)
CODEN: XNBSAV

Library of Congress Catalog Card Number: 81-600108

U.S. GOVERNMENT PRINTING OFFICE
WASHINGTON: 1981

For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, DC 20402
Price $3.00
(Add 25 percent for other than U.S. mailing).

A Survey of Standardization Efforts of Coded Character Sets for Text Processing

Joan E. Knoerdel

ABSTRACT

As the desire to interchange documents among different text processing systems via communications increases, the incompatibilities which exist between those text processing systems become more and more apparent. One such incompatibility is that often the sending device and the receiving device use different bit assignments or coding schemes for their alphanumeric characters, special symbols, and control characters (i.e., different coded character sets).

Considerable work has been done both nationally and internationally to standardize coded character sets. However, the knowledge of such standards efforts is not always widespread. Or, if there is familiarity with the standards efforts, the relationship of those efforts among various standards organizations is frequently not easy to perceive. The objective of this report is to describe the status of those standard coded character sets, with special attention to text processing systems.

The report includes, first of all, a brief description of the major national and international standards organizations which develop code standards. Next, it describes the various code standards according to the following categories: basic code sets for information interchange, methods of augmenting those basic code sets, additional control characters to be used with the basic code sets, and code sets developed specifically for text communications. Finally, the summary of the report discusses a number of limitations which still exist when interchanging information via communicating text processors.

Key words: Code extension techniques; code standards; coded character sets; control functions; information interchange; text communications; text processor.

TABLE OF CONTENTS

1.0 INTRODUCTION

Considerable work has been and is currently being done, both nationally and internationally, in the area of standard coded character sets. However, the relationship of such standards efforts among the standards organizations is not immediately apparent. The objective of this report is to describe the status of those standard coded character sets, with special attention to the area of text processing. Some of the standards discussed herein are in the draft stages or are currently being revised. Therefore, this report discusses only those drafts/revisions developed prior to March 1, 1981.

The report is divided into six sections. This first section defines the key terminology relative to code sets. The second section is devoted to a brief description of the major national and international standards organizations which develop code standards. In the third section, the basic standard code set developed for the exchange of information is introduced. That standard code set includes two types of characters: control and graphic. There are several methods of augmenting those control and graphic characters. Such methods are the subject of section 4. The purpose of the fifth section is to discuss that group of standards which defines control characters that are intended to supplement those in the basic standard code set. In addition, standards efforts are underway to develop a coded character set for the specific application of text communication; section 6 describes that standards effort.

As mentioned above, the remainder of this first section is devoted to defining key terminology which will be germane throughout this report.

A character is a symbol that has a common, constant meaning for some group of people. It might be a letter, a number, or a special symbol. Characters in data communications are represented by groups of bits. The various groups of bits that represent the set of characters that are the "alphabet" of any given system are called a coding system or simply a code.

Codes for representing information in data communications vary in relation to both the number of bits used to define a single character and in the assignment of bit patterns to a particular character. For example, the bit group 1000001 may represent the character A in one coding scheme, but the bit group 11000 may represent the character A in some other code configuration.

One example of a coding system is the American Standard Code for Information Interchange (ASCII). This is a 7-bit code with 128 unique bit combinations. Another example is the Data Interchange Code, which is also a 7-bit code and is primarily used on the slower speed equipment. Other coding schemes which can be used in data communications are the Baudot code, a 5-bit code dating back to the 19th century; the Binary Coded Decimal Code (BCD), a 6-bit code with 64 valid character combinations; or the Extended

Binary Coded Decimal Interchange Code (EBCDIC), an 8-bit code which has 256 valid character combinations.

ASCII is the code which has been adopted as the "standard" code for national and international information interchange. In many instances, ASCII lacks either controls or graphics to sufficiently satisfy the needs of an application. These needs may be satisfied by means of code extension. The primary concept of code extension is to permit the alphabet of a coding scheme to be increased in a manner which is agreeable to the users.

Several techniques exist which can be used to extend the ASCII code set. In order to reduce the risk of conflict between two systems attempting to communicate, it is necessary to use identical code extension techniques. Such techniques are the subject of a set of standards to be described in this report.

In addition to the codes for the basic character set, it is often necessary to manipulate the recording, processing, transmission, or interpretation of that data in some way to make it useful. Such manipulation is accomplished by control functions which are initiated, modified, or terminated by control characters or control codes. These control codes are represented in bit groups as described for the character 'A' in a preceding paragraph. Unlike the character 'A', however, control codes are usually non-printing characters; instead, they serve as instructions or commands for operations such as formatting and editing text or the transmission of data via communications.

The standards which have been developed for information interchange provide a set of 32 control characters which can be augmented through various methods of code extension. The increasing use of text processing systems necessitates that the input and output control features of two-dimensional character-imaging devices (CRT's and printers, buffered and non-buffered) be addressed. A set of standards has been developed which addresses these additional control codes.

2.0 BRIEF DESCRIPTION OF STANDARDS ORGANIZATIONS

This section describes the major national and international standards organizations which develop code standards. The objectives of all the standards bodies are basically the same. These objectives are development of standards, promulgation of standards, coordination of standards development, establishment of standards, and exchange of information.

2.1 International Organization For Standardization (ISO)

The goals of the International Organization for Standardization (ISO) are to promote the development of standards in the world in an effort to facilitate the international exchange of goods and services, and to develop cooperation in the sphere of intellectual, scientific, technological, and economic activity. [22]

To achieve these goals ISO may, among other things:

- Take action to facilitate coordination and unification of national standards and issue necessary recommendations to Member Bodies for this purpose.
- Set up international standards.
- Encourage and facilitate, as occasion demands, the development of new standards having common requirements for use in the national or international sphere.
- Arrange for exchange of information regarding work of its Member Bodies and of its technical committees.
- Cooperate with other international organizations interested in related matters, particularly by undertaking at their request studies relating to standardization projects.

The ISO, which is an international non-governmental organization, has been granted consultative status with the United Nations and many of its agencies.

The ISO membership is composed of those Member Bodies most representative of standardization in their respective countries. Only one such organization in each country may be admitted to ISO.

The major organizational units of ISO are technical committees, which are composed of Member Bodies wishing to take part in the work assigned to individual technical committees. Each technical committee has a secretariat which is a Member Body appointed by the Council. The Council consists of the President and representatives from 14 Member Bodies.

Each technical committee is composed of a number of subcommittees. This is the level at which most of the technical decisions are made, and it is also the level at which much of the technical liaison takes place. Subcommittees are charged with the study of one or several items within the programs of work of the parent technical committee.

The ISO committee charged with character sets and coding standards is TC97/SC2. However, TC97/SC2 receives recommendations from other ISO committees, such as TC97/SC5 (Programming Languages) or TC95/SC15 (Numeric and Alphanumeric Office Machines), concerning coding needs.

2.2 American National Standards Institute (ANSI)

The American National Standards Institute (ANSI) was originally organized in 1918 as the American Engineering Standards Committee. In 1969, this committee was reorganized and the name was changed to ANSI; more recently it has undergone several modifications to its structure. The purpose of the reorganization was to broaden the membership base and encourage user involvement. [22]

Five of the major objectives of the American National Standards Institute are:

- To serve as the national coordinating institution for the development of national standards so as to insure the development of needed standards.
- To provide an independent mechanism for approval and promulgation of voluntary national standards.
- To provide a focal point for industry and Government coordination in the field of standardization.
- To provide the mechanism for managing and coordinating programs of national standards.
- To represent the USA in international standardization organizations of a non-government nature.

ANSI is the official Member Body of the International Organization for Standardization for the U.S. Thus, ANSI provides management, leadership, coordination, and financial as well as administrative support for effective U.S. participation in the international standardization effort. In addition, ANSI helps in governing the ISO through its membership on the ISO Council. In its capacity as Secretariat, ANSI directs the work of many ISO Technical Committees and Subcommittees.

ANSI does not, in itself, develop standards; its only function is to provide the organization through which standards can be developed and approved. The Institute has standards management boards to foster development of standards, a review board to determine that consensus has been reached, and a board for accepting and approving proposed standards.

Within ANSI, the committee X3L2 is the one charged with coding character sets. As with its ISO counterpart (TC97/SC2), X3L2 receives recommendations from other ANSI committees concerning coding needs. For example, X3V1 (formerly X4A12), responsible for office systems standards, and X3J6, responsible for developing standard languages for the processing of text, make inputs to X3L2 on their coding needs.

2.3 National Bureau Of Standards

The Federal Information Processing Standards (FIPS) Publication Series of the National Bureau of Standards is the official publication relating to standards adopted and promulgated under the provisions of Public Law 89-306 (Brooks Act) and under part 6 of Title 15, Code of Federal Regulations. These legislative and executive mandates have given the Secretary of Commerce important responsibilities of improving the utilization and management of computers and automatic data processing in the Federal Government. To carry out the Secretary's responsibilities, the NBS, through its Institute for Computer Sciences and Technology, provides leadership, technical guidance, and coordination of Government efforts in the development of guidelines and standards in these areas [21].

Since it is recognized that information processing standards are being developed within ANSI and ISO, standards developed for Federal requirements are consistent with corresponding ANSI and ISO standards whenever possible. In such cases, NBS is responsible for assuring Federal participation in the development of the standards and for considering them as Federal standards. In the cases where the ANSI or ISO standards do not meet the requirements of the Federal Government, or where ANSI or ISO efforts do not exist, NBS is responsible for initiating independent standards development actions.

The NBS role in the ADP standards program can be summarized as follows:

- providing day-to-day guidance and leadership of an executive branch program to determine requirements for standards and to promote the development and testing of standards for ADP products and services;
- participating in appropriate activities of the National and International voluntary standards organizations;
- monitoring and coordinating all Federal participation in these voluntary activities;
- preparing recommendations for standards to be adopted for Federal implementation;
- monitoring the implementation of Federal standards and assessing their impact on computer services; and
- carrying out the necessary research and analysis in support of the development, implementation, and management of ADP standards.

2.4 European Computer Manufacturers Association (ECMA)

The European Computer Manufacturers Association (ECMA) was officially organized in May, 1961, with a membership consisting of companies which develop, manufacture, and market data processing machines in Europe. The purpose of ECMA as stated in the bylaws is:

- To study and develop, in cooperation with the appropriate national and international organizations, as a scientific endeavor and in the general interest, methods and procedures in order to facilitate and standardize the use of data processing systems.
- To promulgate various standards applicable to the functional design and the use of data processing equipment [22].

ECMA is a non-profit-making organization devoted to no commercial activity.

The committee within ECMA responsible for character sets and coding is TC1. Similar to its international counterparts, TC1 receives inputs from other ECMA committees regarding their coding requirements.

2.5 International Telegraph And Telephone Consultative Committee

The International Telegraph and Telephone Consultative Committee (CCITT) was established in 1957 to examine and make recommendations on questions related to technical, operational, and tariff matters regarding facsimile, telegraph, and telephony [22].

There are five categories of membership (designated A through E) of which only "A" members have voting powers at the Plenary where decisions are made. The United States is an "A" member and is represented by the State Department.

CCITT has liaison status with the International Organization for Standardization (ISO). Recommendation A20, which governs relations with ISO/TC97/SC6 (Data Communications), encourages a spirit of cooperative activity that is complementary rather than competitive, so that there is neither duplication of effort on the part of these two bodies nor unilateral action by either body on matters of mutual interest. ANSI is not eligible for membership in CCITT; however, there is cooperative interaction among these two groups also.

The character set and coding committee within CCITT is Study Group VIII.

3.0 CODE FOR INFORMATION INTERCHANGE

The transfer of data among systems necessitates one code which is understood by all systems. Also, there is a user need for compatibility in the exchange of information. Therefore, a 7-bit coded character set was developed which included both control characters, such as formatting and data communications commands, and graphic characters, such as digits, letters, and special symbols.

One of the considerations affecting the structure of the Code for Information Interchange was the need for an unambiguous code, one in which every code combination had a unique interpretation. The Code for Information Interchange provides this unique representation for 128 code combinations. It is large enough to facilitate all upper and lower case letters of the Latin-based alphabet, the ten decimal digits, some special symbols, and up to

32 control characters. As will be described later, the Code for Information Interchange can be extended to provide an even greater repertoire of characters. The Code for Information Interchange is primarily intended for interchange of information among data processing systems and associated equipment and within message transmission systems.

For convenience, the Code for Information Interchange is often presented in the form of a table, the Basic Code Table, of eight columns and sixteen rows. (See Figure 1.) The columns are referenced from 0-7, containing bits 7, 6, and 5. The rows are referenced from 0-15, containing bits 4, 3, 2, and 1. Columns 0 and 1 are reserved for control characters and columns 2-7 (with the exceptions of positions 2/0 and 7/15 which are always reserved for SPACE and DELETE, respectively) are reserved for special symbols, letters and digits.

Each character has a unique 7-bit representation and can be identified in a variety of ways. One method of identification is by binary representation with bit seven, the high-order bit, and bit one, the low order bit; e.g., the character "K" is 1001011. Representation in a number system with a base greater than two is also common. Thus, this same character could be represented as 113 in base 8 (octal), or 4B in base 16 (hexadecimal), or 75 in base 10 (decimal). Another method is by notation of the column and row. Using this method, the notation for the character "K" is "column 4, row 11", or alternatively as 4/11.

A character allocated to a position in the Basic Code Table may not be placed elsewhere in the table. However, there are options to the table which must be agreed to in advance of information exchange to insure compatibility. A single character must be allocated to each of the positions for which freedom exists or it must be declared to be unused. A code table completed in this way is called a "version."

There is a registered version of the Basic Code Table, the International Reference Version (IRV), which is standardized for international use. It is assumed that the IRV is being used in communications unless the sender and recipient have agreed on another version, such as a national or application-oriented (e.g., word processing) version. Figure 2 shows the IRV.

The responsibility for defining national versions lies with the national standardization bodies. These bodies should exercise the options available and make the required selection. More than one national version can be defined within a country. The different versions should be separately identified, even if they differ only by a single character. Different versions can be registered in the ISO and, therefore, will be internationally recognized in information interchange. Registration of code tables, often called repertoires of characters, is addressed in a subsequent section.

Figure 1

Basic code table

				b$_7$	0	0	0	0	1	1	1	1
				b$_6$	0	0	1	1	0	0	1	1
				b$_5$	0	1	0	1	0	1	0	1
				column	0	1	2	3	4	5	6	7
b$_4$	b$_3$	b$_2$	b$_1$	row								
0	0	0	0	0	NUL	TC$_7$ (DLE)	SP	0	[3]	P	’ [4]	p
0	0	0	1	1	TC$_1$ (SOH)	DC$_1$	!	1	A	Q	a	q
0	0	1	0	2	TC$_2$ (STX)	DC$_2$	" [5]	2	B	R	b	r
0	0	1	1	3	TC$_3$ (ETX)	DC$_3$	£(#) [1]	3	C	S	c	s
0	1	0	0	4	TC$_4$ (EOT)	DC$_4$	$(¤) [3]	4	D	T	d	t
0	1	0	1	5	TC$_5$ (ENQ)	TC$_8$ (NAK)	%	5	E	U	e	u
0	1	1	0	6	TC$_6$ (ACK)	TC$_9$ (SYN)	&	6	F	V	f	v
0	1	1	1	7	BEL	TC$_{10}$ (ETB)	' [6]	7	G	W	g	w
1	0	0	0	8	FE$_0$ (BS)	CAN	(	8	H	X	h	x
1	0	0	1	9	FE$_1$ (HT)	EM	)	9	I	Y	i	y
1	0	1	0	10	FE$_2$ (LF)[1]	SUB	*	:	J	Z	j	z
1	0	1	1	11	FE$_3$ (VT)[1]	ESC	+	;	K	[3]	k	[3]
1	1	0	0	12	FE$_4$ (FF)[1]	IS$_4$ (FS)	, [5]	<	L	[3]	l	[3]
1	1	0	1	13	FE$_5$ (CR)[3]	IS$_3$ (GS)	–	=	M	[3]	m	[3]
1	1	1	0	14	SO	IS$_2$ (RS)	.	>	N	^ [4][5]	n	‾ [4][5]
1	1	1	1	15	SI	IS$_1$ (US)	/	?	O	_	o	DEL

Figure 2

International reference version

				b_7	0	0	0	0	1	1	1	1
				b_6	0	0	1	1	0	0	1	1
				b_5	0	1	0	1	0	1	0	1
				column	0	1	2	3	4	5	6	7
b_4	b_3	b_2	b_1	row								
0	0	0	0	0	NUL	TC_7 (DLE)	SP	0	@	P	’	p
0	0	0	1	1	TC_1 (SOH)	DC_1	!	1	A	Q	a	q
0	0	1	0	2	TC_2 (STX)	DC_2	″	2	B	R	b	r
0	0	1	1	3	TC_3 (ETX)	DC_3	#	3	C	S	c	s
0	1	0	0	4	TC_4 (EQT)	DC_4	¤	4	D	T	d	t
0	1	0	1	5	TC_5 (ENQ)	TC_8 (NAK)	%	5	E	U	e	u
0	1	1	0	6	TC_6 (ACK)	TC_9 (SYN)	&	6	F	V	f	v
0	1	1	1	7	BEL	TC_{10} (ETB)	′	7	G	W	g	w
1	0	0	0	8	FE_0 (BS)	CAN	(	8	H	X	h	x
1	0	0	1	9	FE_1 (HT)	EM	)	9	I	Y	i	y
1	0	1	0	10	FE_2 (LF)	SUB	*	:	J	Z	j	z
1	0	1	1	11	FE_3 (VT)	ESC	+	;	K	[	k	{
1	1	0	0	12	FE_4 (FF)	IS_4 (FS)	,	<	L	\	l	|
1	1	0	1	13	FE_5 (CR)	IS_3 (GS)	-	=	M	]	m	}
1	1	1	0	14	SO	IS_2 (RS)	.	>	N	^	n	-
1	1	1	1	15	SI	IS_1 (US)	/	?	O	_	o	DEL

The Code for Information Interchange includes a legend of control and graphic characters. The legend contains the abbreviation or symbol, the name, and position in the code table. Also, a definition is provided for each of the control characters.

Diacritical marks are accommodated in two ways: introduce accented letters as single characters in the positions so provided in the Code Table, or design some printing symbols for use in the composition of accented letters. To achieve the second option, a sequence of three characters, consisting of a letter, backspace and one of the printing symbols, is needed. An example is: "a" BACKSPACE " " to form the letter "a" with a circumflex. The result is that the is over-printed in the same position as the "a" to form the accented letter. It should be noted that these symbols take on their diacritical significance only when they are preceded or followed by one backspace character.

3.1 Description Of ASCII

The Code for Information Interchange is the basic standard adopted by all the standards organizations. However, as discussed above, each organization has the option of substituting symbols and letters wherever necessary in the locations specified for that purpose to form a national version. In the United States, a national version of the Code for Information Interchange has been defined and is known as American Standard Code for Information Interchange (ASCII).

The control characters are the same in both ASCII and ISO 646. However, the ASCII graphic set differs from the graphic set of the IRV in ISO 646 in the following ways: ASCII has the dollar sign in position 2/4 and the tilde in position 7/14 instead of the currency symbol and the overline, respectively. The circumflex (^), tilde (~), and four punctuation marks (quotation marks, apostrophe, comma, and opening single quotation mark) can be used in conjunction with backspace to form diacritical marks in order to permit the representation of languages other than English [3].

3.2 Available Standards

There are a number of standards available which address 7-bit coded character sets for information interchange. They are:

ISO 646-1973 (7-bit Coded Character Set for Information Processing Interchange)

This international standard contains two code tables: the Basic Code Table and the International Reference Version (IRV). The Basic Code Table consists of a general table with a number of options and explanatory notes in order to provide some flexibility to accommodate special national or application requirements [14].

The IRV, consisting of the Basic Code Table with characters allocated to the optional positions, is used when there is no requirement for a national or application-oriented version. No substitutions can be made in the IRV. The use of a version other than the IRV requires precise agreement among the interested parties.

ANSI X3.4-1977 (Code for Information Interchange)

This standard is identified as the American Standard Code for Information Interchange (ASCII). The differences between the IRV in ISO 646 and ASCII are discussed in section 3.1.

FIPS Publication 1-1 (Code for Information Interchange)

FIPS Pub 1-1 adopts in whole the ANSI Standard Code for Information Interchange X3.4-1977 [20].

ECMA-6 (7 Bit Input/Output Coded Character Set)

ECMA-6 contains only minor differences from ISO 646. Namely, ECMA-6 shows the up-arrow and tilde on its IRV Table in place of the circumflex and overline, respectively, shown in ISO 646 [8].

CCITT Recommendation V.3 (International Alphabet No. 5)

International Alphabet No. 5 is technically identical to ISO 646 [16].

4.0 CODE EXTENSION TECHNIQUES

It has become increasingly apparent that some applications cannot be satisfied by the provisions offered in the Basic Code Table or the IRV. Such applications include scientific or mathematical notation, word processing, and photocomposition. These applications can be satisfied by extending the Basic Code Table or IRV. This section will describe code extension techniques for two environments: 7-bit code environments and 8-bit code environments.

4.1 7-bit Code Environment

ISO 646 and ASCII (both 7-bit codes) have 32 control characters, and 94 graphic characters, plus SPACE and DELETE. The 7-bit code is thought to be near optimum with respect to size; it is small enough that it is not costly to implement or inefficient to transmit, and it is large enough that it accommodates the needs of many users.

However, there will be numerous applications with requirements that are not accommodated by a 7-bit code, or at least not by the specific characters in ISO 646 or ASCII. In order for ISO 646 or ASCII to service more applications, various code extension procedures have been developed.

4.1.1 Techniques Of Code Extension -

There are two techniques of code extension for a 7-bit code: extension by substitution and extension by increasing the repertoire of characters.

4.1.1.1 Extension By Substitution -

This method is simply substituting the needed new character(s) for some of the characters of the Code Table. The characters being substituted into the Table must be of a similarly structured coding system; i.e., they must be from a 7-bit code for 7-bit Code for Information Interchange or from an 8-bit code for 8-bit Code for Information Interchange. An example of this methodology is in applications where there is a need for a paragraph and section symbol but no need for an exclamation mark and a plus sign. The paragraph and section symbols could be substituted in positions 2/1 and 2/11 of the Code Table. Of course, any substitutions necessitate prior agreement between sender and recipient.

Extension by substitution has several advantages. First, it is adaptable to most small and large systems with the use of a keyboard or single key overlay and, possibly, some software and/or hardware adjustments. Secondly, character substitution establishes a basic level of compatibility which can be extended to any number of countries allowing terminals to communicate nationally and internationally. In most cases, a host system is not necessary; thus, communication may proceed point-to-point.

There are three disadvantages to this method of extension. First, some characters already in the set must be sacrificed. Second, the number of control and graphic character substitutions is limited; therefore, it is a poor choice for a non-Latin alphabet containing a large number of characters (e.g., Japanese). Third, the system must be capable of receiving and displaying the full repertoire of characters. Some terminals, such as "dumb" terminals, cannot be altered to accommodate any code extension techniques.

ISO has disregarded further development of code extension by substitution. However, this methodology is under study in the CCITT. It was also the recommended method of extension by ANSI X4A12/WG-3 (Word Processing Alphabet and Code) for use in word processing applications. (This original WG-3 has since dispanded and X4A12 has merged with ANSI X3 to form X3V1.)

4.1.1.2 Extension By Increasing The Repertoire Of Characters -

In this method, the 128 characters of the Code for Information Interchange Code Table or the IRV remain unchanged. However, additional characters can be provided by:

1) additional single controls;
2) additional sets of 32 control characters;
3) additional sets of 94 graphic characters (94 rather than 96 because SPACE and DEL remain in position 2/0 and 7/15, repectively, in all graphic character sets);
4) additional sets of more than 94 graphic characters each represented by more than one byte.

Examples of each method of providing additional characters follow. 1) An example of an additional single control character is Identify Graphic Sub-repertoire (IGS), which can be used in a data stream to instruct the recipient to change the printing element on the receiving device. 2) If more than one or two additional control characters are needed for a particular application, a set of up to 32 additional control characters can be identified and introduced. An example is the set of special controls needed in typesetting. 3) Additional sets of graphic characters can also be used, such as the French graphic set. Since the French language requires extensive use of accented letters, it is convenient to incorporate them into a set in order to eliminate the need to produce them in multiple steps. 4) The code of the Japanese graphic character set for information interchange is an example of a multi-byte set. The first byte is the row position and the second byte is the column position of the graphic character.

Sets of control characters have been named C0 and C1; sets of graphic characters have been named G0 and G1. Generally, C0 and G0 sets are those sets which are used frequently; while C1 and G1 sets are those sets which are used less frequently. It is not a requirement that all the positions in C0, C1, G0, and G1 sets be filled.

Often, it is desirable to use these control and graphic sets in combination. In fact, many applications require such combination. For instance, in a text communications application the C0 and G0 sets can consist of the Code for Information Interchange controls and graphics, respectively. The C1 set can consist of additional controls, such as super- and sub-script, and a code extension control function (e.g., Control Sequence Introducer). The G1 set can be composed of accented letters and non-Code for Information Interchange symbols, such as paragraph symbol, section symbol, product dot, and the quarter and half symbols.

The advantages of extension by increasing the charater repertoire include the ability to support non-Latin alphabets with a large number of characters. Additionally, full repertoires of graphic and control character sets can be accommodated for different applications, such as word processing control characters and graphic characters or font control codes. Another advantage is that transfer of data between a 7-bit environment and an 8-bit environment can be accomplished.

However, there are some disadvantages of increasing the repertoire of characters. One is the possible confusion for the user as the number of character and control sets increases. Further, this method is geared towards larger, more sophisticated equipment; therefore, much consideration must be given to the fact that this methodology eliminates a large percent of the equipment currently in use.

4.1.2 Methods Of Accomplishing Code Extension -

There are two means of accomplishing code extension: extension by means of shift-in/shift-out and extension by means of escape sequences.

4.1.2.1 Extension By Means Of Shift-in/Shift-out -

The shift-in/shift-out characters are used exclusively for the extension of graphics, and, therefore, do not impact the control characters.

The shift-out (SO) control character shifts from a G0 set to an alternative set of 94 characters, the G1 set. The shift-in (SI) control character is used to return to the G0 set from the G1 set. While in the SO state (i.e., working in the G1 set), the SO character has no effect, and, while in the SI state (i.e., working in the G0 set), the SI character has no effect.

When using an alternative set (G1 set), it is not necessary to assign a graphic character to all positions of that additional set; in fact there is no requirement that all the graphic characters be different from those of the G0 set. However, it is required that SPACE and DEL remain in positions 2/0 and 7/15, respectively, in any new G0 or G1 set. Also, any characters to be used in a planned escape sequence (discussed in the following section) must remain in the same positions in any of the G0 or G1 sets.

If an application requires more than one G0 set and G1 set, it is necessary to identify each set with a unique group of characters which begins with the escape character (1/11) and is called an escape sequence. When multiple G0 sets and/or G1 sets exist within a system, it is possible to move from one G0 set to another G0 set or to move from a G0 set to a G1 set and return. In moving from one G0 set to another G0 set, the second G0 set is invoked immediately following the escape sequence identifying it. In moving from a G0 set to a G1 set, the desired G1 set is identified by its identifying escape sequence but not invoked until the SO character occurs. The reverse is also true; i.e., when working within a G1 set the designated G0 set is not invoked until the SI occurs. Once the G0 and G1 sets are designated, they remain in effect until one or both are changed by another escape sequence.

4.1.2.2 Extension By Means Of Escape Sequences -

Escape sequences have five uses. First, they can represent single control functions in addition to the 32 control functions provided in Code for Information Interchange. Second, escape sequences can provide additional sets of 32 control functions. These single control functions and sets of control functions cannot include transmission control functions. The third use of escape sequences is to identify additional sets of 94 graphic characters not already in ASCII. A fourth use of an escape sequence is to indicate that a 7-bit combination will have a use other than its standard use. And, fifth, escape sequences are used to designate coded character sets with a number of bits other than 7.

An escape sequence consists of two or more of the following components: Escape (ESC), Intermediate character(s), and a Final character, in that order. ESC is an ASCII control character in position 1/11 of the code table. Intermediate characters, denoted by the symbol (I), are the 16 characters of column 2 of the 7-bit code table. Final characters, denoted by the symbol (F), are any one of the 79 characters of columns 3 to 7 of the 7-bit code table, excluding DEL. The control characters in columns 0 and 1 are prohibited from use as Intermediate and Final characters.

A two-character escape sequence takes the form:

ESC (F)

Two-character escape sequences get their meanings from their Final characters and are used to represent single additional control characters. As seen in Figure 3, the 7-bit code table is divided into three "fields": the field Fp consists of column 3; the field Fe consists of columns 4 and 5; and the field Fs consists of columns 6 and 7. These fields determine the meaning of (F). A Final character from field Fp indicates a single additional control character which does not have a standardized meaning but is used for private use. A Final character from field Fe indicates an individual control character of an additional standardized set of 32 control characters (i.e., an individual control character from the standardized C1 set). A Final character from field Fs indicates a single additional standardized control character.

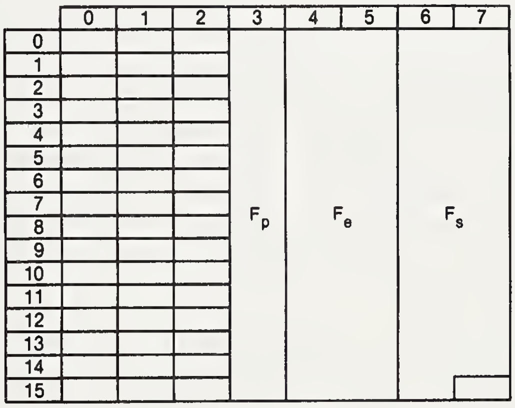

Figure 3

A three-character escape sequence takes the form:

ESC (I) (F)

The Intermediate characters consist of the characters in column 2. Final characters may belong to one of 2 fields: Fp, as defined for the two-character escape sequence, and Ft, which consists of the characters in columns 4 to 7 of the code table. Fp is reserved for private use, while Ft is standardized. Figure 4 demonstrates the fields for the 3-character escape sequences.

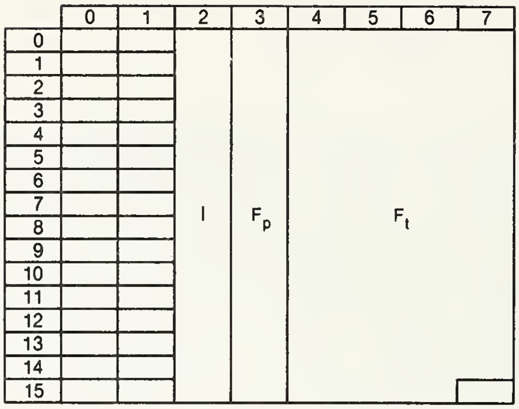

Figure 4

The meaning of the three-character escape sequence is determined by its Intermediate character and Final character. Depending on the Intermediate character, three-character escape sequences can: represent a single additional control character; designate and invoke a C0 set or C1 set of 32 control characters; designate sets of 94 graphic characters which will be used as a G1 or G0 set; designate and invoke a code that requires special interpretation; or, designate sets of graphic characters that are represented by two or more bytes each corresponding to a bit combination in columns 2 to 7.

A single additional control character is represented by the structure ESC 2/3 (F). The structure which designates and invokes a C1 set is ESC 2/2 (F); and, the structure which designates and invokes a C0 set is ESC 2/1 (F).

A set of 94 graphic characters which will be used as a G0 set is designated by either ESC 2/8 (F) or ESC 2/12 (F). The designated G0 set is invoked by SI. A set of 94 graphic characters to be used as a G1 set is designated by either ESC 2/9 (F) or ESC 2/13 (F). The designated G1 set is invoked by SO.

A code requiring special interpretation is designated and invoked by the sequence ESC 2/5 (F). Examples of such a code are: one with a number of bits other than 7, excluding those 8-bit codes structured in accordance with this standard; and, a 7-bit code whose characteristics differ from those in this standard.

ESC 2/4 (F) designates sets of graphic characters that are represented by two or more bytes, each corresponding to a bit combination in columns 2 to 7. These sets are regarded as G0 sets. Therefore, if the multi-byte set is identified while in the G1 set, it must be invoked by SI.

In the case of escape sequences having four or more characters, the meanings of the first Intermediate character and the Final character are identical to those defined in the three-character sequences. The remaining Intermediate character(s) permit additional subcategories within the category defined by the first Intermediate character.

The structure and use of escape sequences are fully documented in the set of standards discussed in this section. However, the specific meanings assigned to individual escape sequences are not defined. Instead, their meanings are established by an ISO registration procedure described in section 4.4.

4.2 8-bit Code Environment

An 8-bit code inherently provides for twice the number of characters as a 7-bit code. An example of an 8-bit code is the 8-bit ASCII code. The standard for information interchange defines a 7-bit code. In addition, the standard for code extension defines an 8-bit version of the 7-bit ASCII code. When the ASCII 7-bit code is extended to an 8-bit code, the eighth bit is treated as a high-order bit and is set to zero (0) for the 128 ASCII characters and is set to one (1) for the additional 128 characters. For example, the bit configuration for the letter K in 7-bit ASCII is 1001011; while in 8-bit ASCII, it is 01001011. The characters assigned to columns 00-07 of the 8-bit code table are identical to those in the 7-bit ASCII code table. The characters assigned to columns 08-15 of the 8-bit code table have been identified in several application areas, such as for bibliographic information interchange on magnetic tape (ANSI Z39.2-1971).

Another example of an 8-bit code is the Extended Binary Coded Decimal Interchange Code (EBCDIC). The table is arranged so that all control characters are assigned to columns 0-3, special symbols to columns 4-7, lower case alphabetic characters to columns 8-10, upper case alphabetic characters to columns 12-14, and digits to column 15. The EBCDIC code table is shown in Figure 5 below.

It should be noted that any reference to an 8-bit code in this report is to 8-bit ASCII.

4.2.1 Introduction To 8-bit ASCII -

The 8-bit code table is represented as an array of 16 columns, numbered 00 to 15, and 16 rows, numbered 0 to 15. As can be seen in Figure 6, the table is divided into four sections: a C0 set, a G0 set, a C1 set, and a G1 set. The characters of the C0 set and G0 set are identical to the characters in 7-bit ASCII. However, the appearance of the code table varies slightly: the

EBCDIC Code Table

				B8	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
				B7	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
				B6	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
				B5	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
B4 ↓	B3 ↓	B2 ↓	B1 ↓	HEX-0 HEX-1	0	1	2	3	4	5	6	7	8	9	A	B	C	0	E	F
0	0	0	0	0	NUL	OLE	OS		SP	&										0
0	0	0	1	1	SOH	SBA	SOS				/		a	j			A	J		1
0	0	1	0	2	STX	SUA	FS	SYN					b	k	s		B	K	S	2
0	0	1	1	3	ETX	IC							c	l	t		C	L	T	3
0	1	0	0	4	PF	RES	BYP	PN					d	m	u		0	M	U	4
0	1	0	1	5	PT	NL	LF	RS					e	n	v		E	N	V	5
0	1	1	0	6	LC		ETB	UC					1	o	w		F	0	W	6
0	1	1	1	7	OEL	IL	ESC	EOT					g	p	x		G	P	X	7
1	0	0	0	B		CAN							h	q	y		H	0	Y	B
1	0	0	1	9		EM				·			I	r	z		I	R	Z	9
1	0	1	0	A	SMM	CC	SM		¢	!	¦	:								
1	0	1	1	B	VT				.	$	,	#								
1	1	0	0	C	FF	OUP		RA	<	•	%	@								
1	1	0	1	0	CR	SF	ENO	NAK	(	)	–	'								
1	1	1	0	E	SO	FM	ACK		+	;	>	=								
1	1	1	1	F	SI	ITB	BEL	SUB	|	¬	?	"								

Figure 5

column numbers have been preceded with a "0" (00, 01, 02, etc.); and, as mentioned above, a high-order bit of zero (0) has been added to the bit combination for each ASCII character.

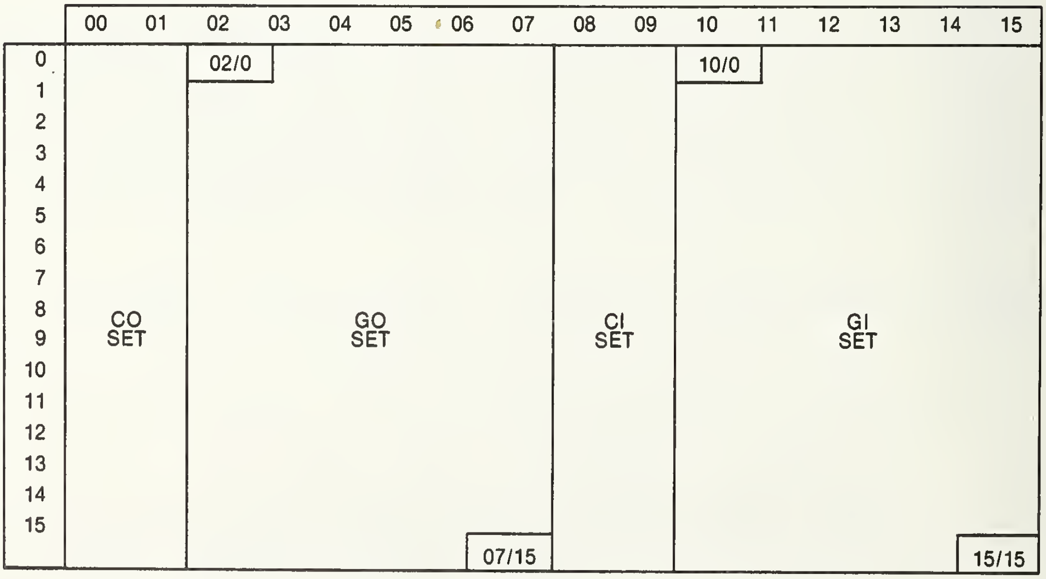

Figure 6
The 8-Bit Code Table

The C1 set and G1 set have column numbers 08-09 and 10-15, respectively. The high-order bit for all C1 and G1 characters is 1. These 128 characters are restricted in the following ways: columns 08 and 09 are provided for an additional set of 32 control characters; no transmission control characters can be included; columns 10 to 15 provide for a set of 94 additional graphic characters, excluding positions 10/0 and 15/15.

4.2.2 Code Extension In An 8-bit Code -

The construction of escape sequences in an 8-bit code is exactly the same as in the 7-bit code: ESC (F) or ESC (I) (F), and are composed of characters of the C0 and G0 sets only. All characters in columns 08-15 (the C1 and G1 sets) are excluded from use in escape sequences. The meanings of the escape sequences in the 8-bit code are identical to those defined for the 7-bit code.

The structure of the escape sequences is identical to the 7-bit environment with one exception. In a two-character sequence, the use of the ESC Fe sequence to identify the additional set of control characters is unnecessary because in the 8-bit environment the C1 set is immediately accessible to the user. However, should an ESC Fe sequence occur, its meaning is the same as in the 7-bit environment.

4.3 Retention Of The 7-bit Code In The 8-bit Environment

At times, it is desirable to retain information in 7-bit form while in an 8-bit environment. In that situation, bits b7-b1 in 7-bit ASCII are represented in 8-bit ASCII as a7-a1, and the high order bit (a8) is set to 0. In transferring between 7-bit and 8-bit codes, indication that 7-bit or 8-bit coded data follows is achieved by the proper announcing escape sequence. For example, the sequence ESC 2/0 4/3 or ESC 2/0 4/4 indicates that 8-bit coded data follows, and the sequence ESC 2/0 4/1 or ESC 2/0 4/2 announces that 7-bit data follows.

4.4 Registration Of Code Tables

In order to promote compatibility in information interchange, the existence of application-oriented character sets must be known to all potential users. Such common knowledge also eliminates duplication of effort in developing application-oriented character sets. In addition to having knowledge of the character sets, users must also be aware of the escape sequences identifying those character sets.

The ISO has developed a standard which outlines the procedure to be followed in preparing and maintaining a register of character sets and their identifying escape sequences. That standard is ISO 2375, Data Processing - Procedure for Registration of Escape Sequences [13].

According to ISO 2375, the initial step in registering a set is to make application to a sponsoring authority. A sponsoring authority is defined as any ISO technical committee, subcommittee, or working group, any Member Body of ISO, or any international organization having liaison status with ISO.

The proposal is submitted in a standard form, the layout of which is available from the registration authority. The registration authority is a member of ISO/TC97/SC2 and is appointed by that committee. Currently, the Association Francaise de Normalisation (AFNOR) is serving as the registration authority.

As the registration authority, AFNOR assigns an identifying escape sequence to a proposed character set. The escape sequence and character set are then voted on by the members of ISO/TC97/SC2. If approved, AFNOR informs all ISO Member Bodies and ISO liaison organizations of the escape sequence and character set.

4.5 Available Standards

The standards on code extension adopted by the various standards organizations are technically identical. Those standards are:

ISO 2022-1973 (Code extension techniques for use with the ISO 7-bit coded character set) [10]

ANSI X3.41-1974 (Code extension techniques for use with the 7-bit coded character set of ANS code for information interchange) [2]

FIPS PUB 35 (Code extension techniques in 7 or 8 bits) [19]

ECMA-35 (Extension of the 7-bit coded character set) [7]

5.0 ADDITIONAL CONTROL FUNCTIONS

With the proliferation of character-imaging devices (e.g., word processing equipment, printers, and CRT display devices) in the marketplace and the desire to communicate among those devices, it has become increasingly apparent that the control functions defined by ANSI X3.4-1977 and ISO 646 are insufficient. A set of additional control functions, which can be used in combination with the C0 set defined by ANSI X3.4 and ISO 646, has been developed.

These additional control functions facilitate diverse information interchange in the following applications:

- Interactive terminals of the CRT type;
- Interactive terminals of the printer type;
- Line printers;
- Microfilm printers;
- Software usage;
- Form filling;
- Composition imaging, e.g., typesetting;
- Word processing;
- Input/output devices with auxiliary devices; and
- Buffered and non-buffered devices.

Additional control functions provide flexibility in the formatting and editing capabilities available to the user along with the accessibility of a variety of character repertoires through code extension. It is possible to implement only the additional control functions which are needed for the application. Also, the incompatibility in control codes between vendors of character-imaging devices currently being designed (e.g., word processing) can be reduced or, possibly, eliminated with the implementation of some of these additional control functions. However, this is generally not true of character-imaging devices now in the marketplace because much of that equipment cannot accommodate additional control functions. Even if it were possible to alter the existing equipment, the cost of doing so might not be justified.

Each of the additional control functions belongs to one of the following categories:

- control functions which are independent,
- control functions represented by ESC Fs sequences,
- control functions represented by control strings, or
- control functions represented by control sequences.

It is important to retain the concept that control function implies a string of bits taken together as a whole regardless of the number of characters used in their representation.

5.1 Independent Control Functions

Currently, there are 25 functions which are defined as independent control functions (i.e., controls that do not have an introducer). These control functions are sometimes referred to as elements of the C1 set. They include controls which:

1) affect the interpretation of the data stream; for example, those which introduce additional single graphic characters (i.e., Single Shift 2 (SS2) and Single Shift 3 (SS3)); introduce control sequences (i.e., Control Sequence Introducer (CSI)); or delimit control strings

for software and hardware/system strings (i.e., Application Program Command (APC), Device Control String (DCS), and String Terminator (ST));

2) solicit a response from the receiving device; for example, those which permit priority interchange, such as Message Waiting (MW); or,

3) are represented by a single bit combination to facilitate their storage in a buffer; for example, area delimiters, such as End of Protected Area (EPA) and Start of Protected Area (SPA); or, format effectors, such as Partial Line Up (PLU) and Vertical Tab Set (VTS).

The independent control functions which are of particular interest in the word processing environment are those which introduce control functions or sequences and those which are format effectors because their use by word processor equipment vendors can provide compatibility of formatting and editing controls among unlike equipment.

In accordance with the code extension techniques defined in ANSI X3.41 and ISO 2022, each independent control function can be represented by a 2-character escape sequence in a 7-bit code, i.e., ESC Fe, or by a single bit combination from column 08 or 09 in an 8-bit code. For example, CSI is 1/11 5/11 in 7 bits and is 9/11 in 8 bits.

5.2 ESC Fs Sequences

These escape sequences are single additional controls which are defined in the code extension standard as 2-character sequences in the form ESC Fs, where Fs is a bit combination from positions 6/0 to 7/14.

The following four control functions are represented by ESC Fs sequences:

NAME (Abbrev.)	FINAL CHAR
Disable Manual Input (DMI)	6/0
Interrupt (INT)	6/1
Enable Manual Input (EMI)	6/2
Reset to Initial State (RIS)	6/3

ESC Fs sequences are available independent of the control set which has been designated as the current C0 or C1 set. These are functions which are important no matter what control set is being used.

5.3 Control Strings

A control string is a delimited string of characters which may occur in the data stream as a logical entity for control purposes. Two families of strings are established: those which are used in a software application and are provided for communication by a human with a system, and those which are used in a hardware application and are provided for passing information within portions of the system. A control string consists of an opening delimiter (an independent control function), followed by a string of graphic characters and spaces, and the closing delimiter, STRING TERMINATOR (ST).

The software control strings are divided into 3 areas: the privacy discipline, the operating system control program, and the actual application program. The corresponding opening delimiters are Privacy Message (PM), Operating System Command (OSC), and the Application Program Command (APC). A privacy message can be transmitted by enclosing it within the delimiters PM and ST. The interpretation of the privacy message is subject to the individual privacy and security methods in effect. Similarly, a user can transmit a command to an operating system (e.g., print a file) by enclosng the command within the delimiters OSC and ST. An example of the use of an Application Program Command string is the interjection of application program commands into a data stream or into a data file being processed by the application program.

There are many devices whose control is determined by the receipt of a command string at their interface. These capabilities and/or requirements are identified by a Device Control String (DCS). DCS strings take the form: DCS Gs...Gs ST, where Gs is a graphic character from 2/0 to 7/14 appearing in a string. Examples of applications for DCS's are program loading, configuration control, mode control, and diagnostics.

5.4 Control Sequence Functions

The Control Sequence Introducer (CSI) introduces sequences of bit combinations similar to the ESC in escape sequences. Since CSI serves only as an introducer, it must be accompanied by additional information which identifies the specific function to be performed and any needed parameters. Control sequence functions include editing controls, formatting controls, cursor controls, and scrolling controls.

The format for a control sequence function is as follows:

CSI P...P I...I F

where:

CSI is an independent control function which is used as an introducer. It is represented by the two-character escape Fs sequence 1/11 5/11 in a 7-bit code and by 09/11 in an 8-bit code.

P...P is the parameter string, taken from column 3 of the ASCII code table. Parameter strings are divided into 2 types: numeric and selective (representing a character string). They provide detailed information within the desired function, such as the number of lines to be deleted, the type of justification desired (e.g., right flush, word fill), movement of the cursor relative to its current position, and the number of character positions to move to the left or right. Parameter values are not required in all control sequence functions.

I...I represents the intermediate characters from column 2 of the ASCII code table. These characters are used to expand the number of control sequence functions which can be defined. For example, control codes in the word processing environment could be expanded by using new intermediate characters in existing control sequence functions.

F is the function defining character. If F is from columns 4, 5, or 6, it is standardized or reserved for standardization; if F is 7/0 through 7/14, it is available for private or experimental use. Figure 7 defines the final bit combinations allocated to control sequences without intermediates. However, they may require parameter value(s). Figure 8 defines the final bit combinations allocated to control sequences with 2/0 as a single intermediate. They may also require parameter value(s).

Row No.	Column No.		
	4	5	6
0	ICH	DCH	HPA
1	CUU	SEE	HPR
2	CUD	CPR	REP
3	CUF	SU	DA
4	CUB	SD	VPA
5	CNL	NP	VPR
6	CPL	PP	HVP
7	CHA	CTC	TBC
8	CUP	ECH	SM
9	CHT	CVT	MC
10	ED	CBT	
11	EL		
12	IL		RM
13	DL		SGR
14	EF		DSR
15	EA		DAQ

Figure 7

ALLOCATION OF FINAL BIT COMBINATIONS TO CONTROL SEQUENCES WITHOUT INTERMEDIATES

Row No.	Column No.		
	4	5	6
0	SL		
1	SR		
2	GSM		
3	GSS		
4	FNT		
5	TSS		
6	JFY		
7	SPI		
8	QUAD		
9	SSU		
10			
11			
12			
13			
14			
15			

Figure 8

ALLOCATION OF FINAL BIT COMBINATIONS TO CONTROL SEQUENCES WITH 2/0 AS A SINGLE INTERMEDIATE

5.4.1 Parameter Values -

As mentioned above, a parameter string consists of bit combinations from column 3 of the ASCII Code Table. If the first bit combination is 3/0 - 3/9, it represents the corresponding decimal digit 0-9. However, 3/0 as a leading bit combination is insignificant and may be omitted. Parameter values are separated by a parameter separator, 3/11 (semicolon). The bit combination 3/10 (colon) is reserved for future standardization as an additional parameter separator. An empty parameter string, e.g., CSI ;; ST, or a parameter string of only 3/0 represents a default value of the control function. Thus, a parameter string contains one or more parameter values which are groups of one or more bit combinations terminated by a parameter separator.

If the first bit combination is in the range 3/12-3/15, the parameter string is for private use and is not specified here. These bit combinations should not be used in a parameter string as described in the previous paragraphs.

An example of a control sequence function with numeric parameters is Font Selection (FNT) which designates a stylistic font. The format is:

CSI Pn 3/11 Pn I F

The first parameter specifies the primary or alternative font to be used. The second parameter identifies the stylistic font to be designated. The default stylistic font is implementation defined. The parameters may be 3/0 (default font) and 3/1 to 3/9 representing the first alternative font to the ninth alternative font. The designated font is invoked by subsequent Select Graphic Rendition (SGR) control. Up to nine alternative fonts can be designated with this combination of control strings. At any point in the data stream, the font of the following data can be changed with the appropriate SGR control sequence to one of these designated fonts. In this example, the intermediate character (I) for FNT would be bit combination 2/0, and the final character (F) would be bit combination 4/4.

In control sequence functions with selective parameters, each parameter value represents one selective parameter. These values, even though expressed by digits, are not quantitative. Each corresponds to one of the actions the control function can perform.

An example of a control sequence function with a selective parameter is the format effector Justify (JFY). This format effector indicates the beginning of a string of character positions, the contents of which are to be justified according to the layout specified by a parameter value. The parameter values are:

0 end of justification,
1 word fill,

2 interword space,
3 letter space,
4 hyphenation,
5 flush to left margin,
6 center between margins,
7 flush to right margin, and
8 Italian hyphenation.

The end of the string to be justified is indicated by the next occurrence of JFY in the data stream.

Figure 9 shows several examples of parameter strings. The character form is the form input by the operator, while the bit combination form is the position of each character on the ASCII Code Table.

Character Form	Bit-Combination Form	Explanation
7	3/7	A parameter value of 7
98	3/9 3/8	A parameter value of 98
4;2	3/4 3/11 3/2	Two parameters with values 4 and 2
1;;4	3/1 3/11 3/11 3/4	Three parameters with the first value 1, the middle value the default, and the third value 4

Figure 9
Examples of Parameter Strings

5.4.2 Mode Controls -

There are three control sequences which alter the meaning of subsequent control functions. They are Set Mode (SM), Reset Mode (RM), and Set Editing Extent Mode (SEM). In the case of SM and RM, the selective parameter(s) of the control sequences identify which mode is to be set or reset. For SEM, the selective parameter determines how much text will be affected by the control sequence. Modes address the way in which a character-imaging device transmits, receives, processes, or presents data.

The methods of implementing these modes can vary: some or all of the modes may have fixed values, incapable of being set or reset explicitly; some modes may be software or hardware implemented; and others may be established explicitly within the data stream or by previous agreement between sender and receiver.

The selective parameters identifying these modes are divided into four classes:

1) Modes which apply to a device transmitting a data stream or transferring data to an auxiliary device. They are:

FETM	Format Effector Transfer Mode
GATM	Guarded Area Transfer Mode
MATM	Multiple Area Transfer Mode
SATM	Selected Area Transfer Mode
SRTM	Status Reporting Transfer Mode
TTM	Transfer Termination Mode

2) Modes which apply locally to the device and do not affect the device when sending or receiving. They are:

CRM	Control Representation Mode
KAM	Keyboard Action Mode
SRM	Send-Receive Mode

3) Modes which apply locally to a device receiving a data stream and transferring data from an auxiliary device. They are:

EBM	Editing Boundary Mode
ERM	Erasure Mode
FEAM	Format Effector Action Mode
HEM	Horizontal Editing Mode
IRM	Insertion-Replacement Mode
PUM	Positioning Unit Mode
SEM	Select Editing Extent Mode
TSM	Tabulation Stop Mode
VEM	Vertical Editing Mode

4) A mode which applies when a device is transmitting or receiving.

LNM	Line Feed New Line Mode

Four of the modes are designed to interact with each other. They are Guarded Area Transfer, Multiple Area Transfer, Selected Area Transfer, and Transfer Termination. These modes have a combined effect on the format of a transmitted data stream or in the format of a data stream transferred to an auxiliary input/output device.

An example of mode interaction is the following. If the Transfer Termination Mode is reset (i.e., the data up to, but not including, the cursor position is selected for transmission), the Selected Area Transfer Mode is reset (i.e., only the selected data is to be transmitted rather than the entire buffer), and the Multiple Area Transfer Mode is set (i.e., all selected areas of data are to be transmitted), then the contents of any selected areas, up to but not including the cursor position, are marked as the data desired for transmission. The actual initiation of transmission of the data is performed by a data communication or input/output interface control procedure.

5.5 Other Considerations

5.5.1 Format Effectors And Editor Functions -

Two classes of control functions have an action on the layout or positioning of information in character-imaging devices. They are format effectors and editor functions.

Format effectors are intended to remain in the data stream for subsequent execution. Some or all format effectors can be executed immediately upon receipt of the data stream or stored in the data stream for later execution. The time of execution is defined by a selective parameter of the Reset Mode and Set Mode control sequences discussed above, i.e., Format Effector Action Mode (FEAM). In the reset state (RM) of FEAM, format effectors are performed immediately when received in a data stream. However, they are not performed immediately if FEAM is in the set state (SM).

Format effectors describe how the originator wishes the data to be formatted. Therefore, if the receiving device does not store the format effectors with the text (regardless of whether or not they have been executed), they should be stored in a separate file by that receiving device so that they can be reinstated into the text if subsequent transmission is desired.

Editor functions are supplementary control functions whose main purpose is to edit, alter, or transpose the visual arrangement of data previously entered. Therefore, they are usually not stored with the data stream. An example of an editor function which is transmitted in the data stream is Cursor Backwards (CUB) when insertion of information into previously transmitted data is desired.

5.5.2 Selected And Qualified Areas -

A number of the modes interact with a selected area or a qualified area. A selected area is a string of character positions, the contents of which may be selected for transmission in the form of a data stream or for transfer to an auxiliary input/output device. The beginning of the selected area is established by Start of Selected Area (SSA), and the end of the selected area is established by End of Selected Area (ESA).

A qualified area is a string of character positions with which certain characteristics are associated, such as protection against manual alteration or restriction of the set of characters which can be entered (e.g., numeric or alphabetic characters only). The delimiter of qualified areas is DEFINE AREA QUALIFICATION (DAQ).

Two special cases of qualified areas are protected area and guarded area. A protected area is one in which the string of character positions is protected against manual alteration. A guarded area is a protected area which is excluded from transmission in a data stream and from transfer to an auxiliary input/output device.

5.6 Available Standards

The standards addressing additional control functions are as follows:

ISO DIS 6429 (Additional Control Functions for Character-Imaging Devices).

> This standard is presently a draft international standard (DIS) and is being circulated for comment and approval. Any comments which are received will be incorporated into the next revision of the DIS [9].

ANSI X3.64-1979 (Additional Controls for Use with ASCII).

> This standard is a subset of ISO DIS 6429, except for the use of the parameter value LNM with Set Mode and Reset Mode. Revisions of X3.64 will attempt to incorporate those portions of ISO DIS 6429 which presently are not covered [1].

ECMA-48 (Additional Control Functions for Character-Imaging Devices).

> ECMA-48 and ANSI X3.64 represent a coordinated effort to develop a single technical standard in the U.S. and Europe [6].

6.0 ADDITIONAL PROPOSED STANDARDS

6.1 Revision Of ISO 2022

Revisions to ISO 2022 [11], Code Extension Techniques for Use with ISO 7-Bit Coded Character Set, are currently being considered. These revisions expand the current standard by introducing the G2 and G3 sets and two additional extension control characters, namely, Single-Shift 2 (SS2) and Single-Shift 3 (SS3).

The SS2 and SS3 characters are used exclusively for extension of graphics. SS2 invokes one character from the last designated G2 set (designated by an ESC sequence). SS3 invokes one character from the last designated G3 set. Both SS2 and SS3 are non-locking extension characters, which means they affect only the next single bit combination. SS2 indicates that the next bit combination is to be the corresponding bit combination of the G2 set; and, SS3 indicates the corresponding bit combination of the G3 set. The current shift-in or shift-out status is not affected by single-shift characters.

If an application requires the use of only one G2 and/or G3 set, that set is identified by its assigned ESC sequence, if it has one, or by agreement between the interchanging parties. In applications requiring the co-existence of two or more G2 and/or G3 sets, the set to be used next is designated by its identifying ESC sequence.

The structure of the ESC sequences is identical to those in ISO 2022, ESC F or ESC I F. Either ESC 2/10 F or ESC 2/14 F designates 94 graphic characters to be used as a G2 set; a character from the designated set is invoked by SS2 or a locking shift 2. Either ESC 2/11 F or ESC 2/15 F designates 94 graphic characters to be used as a G3 set; a character from the designated set is invoked by SS3 or a locking shift 3.

A complete code can be designated and invoked by use of ESC 2/5 F. Depending on the final character, the broad classifications of codes are:

1) a private code with any number of bits,
2) a code of less than 7 bits,
3) a code of 7 bits,
4) a code of 8 bits, and
5) a code of more than 8 bits.

Extension of the graphic set in an 8-bit code by means of SS2 and SS3 is the same as in a 7-bit code. The bit combinations following SS2 or SS3 are characters from columns 02 to 07, except Space and Delete. All characters in columns 08-15 are excluded from assignment following SS characters.

The changes proposed in this draft standard permit the designation of 376 graphic characters concurrently. Also, the use of ESC 2/5 F sequences are no longer limited to codes requiring special interpretation, but now permit a single escape sequence to designate and invoke a complete 7-bit or 8-bit set. These changes greatly expand the number of characters which can be accessed and the number of repertoires available to users.

The revision of ISO 2022 as described herein could greatly simplify the formation of accented letters by placing the diacritical marks in a G2 or G3 set, which can be used as either a locking set or a nonlocking set. When used as a nonlocking set, no keystroke is required to return to the G0 or G1 set. It would only be necessary to invoke the designated G2 or G3 set with the proper single-shift character and hit the desired diacritical mark with an immediate return to the G0 or G1 set.

If frequent use of an additional set is planned, a combination of G0 and G1 sets is more desirable. That way, once the alternate set is identified, its invocation requires only a Shift-out or Shift-in character in 7 bits and requires no shifting at all in 8 bits. If those sets are invoked as locking sets, they remain in effect until explicitly changed.

A major issue to be considered is that this standard seems to be geared toward larger, more sophisticated equipment. A machine must be able to load new code tables that are registered; most small equipment does not have that capability.

6.2 ISO DP 6937 (Coded Character Sets For Text Communication)

DP 6937 is currently a working draft of an International Standard entitled Coded Character Sets for Text Communication [12]. It consists of three parts as follows: Part 1 - General Introduction, Part 2 - Latin Alphabetic and Non-alphabetic Graphic Characters, and Part 3 - Control Functions for Page-Image Format. The working committee producing this draft proposal is ISO/TC97/SC2/WG4.

The specific purpose of each part is as follows:

> Part 1 defines terms, describes the general field of text communication, and specifies conformance requirements for equipment.
>
> Part 2 defines the repertoire of Latin alphabetic and non-alphabetic graphic characters and its coding, together with the rules for the definition of graphic character subrepertoires.
>
> Part 3 defines the repertoire of control function for communication of text in page-image format and the coded representation for those control functions.

It is the intention of the working committee to add other parts to this standard for inclusion of graphic characters for non-Latin alphabets and pictorial data, and control functions for display-based information retrieval systems.

6.2.1 Part 1 -

Part 1 of ISO 6937 defines terms applicable to all parts of this standard; describes the general field of text communication; describes the general structure of the text communication code; specifies conformance requirements for text communication terminal equipment which relate to the coded character set; and, specifies fall-back presentation requirements for output devices of text communication terminal equipment. Fall-back is a technique whereby, if a terminal device is unable to present a graphic character exactly or to obey a control function exactly, an approximate or equivalent form may be used. For example, a $ may be presented as S overstruck by / or |.

ISO 6937 applies to the communication of text in "page-image format" using public communication networks, private communication networks, and interchangeable media such as magnetic tapes and disks. Page image format is defined as:

> a representation of data for text communication which is formatted by the sender for presentation by the receiver and which is not intended to be subjected to text processing operations by the receiver. (NOTE: In this context, the term "page" is used in a general sense including any two-dimensional presentation medium such as a display screen.)

The kinds of text to which the character repertoire defined in this standard is applicable are those capable of being produced on equipment, such as office typewriters and computer peripheral devices, which would include communicating word processors. Examples of those kinds of text are:

1) business correspondence;
2) professional correspondence and reports;
3) legal and contractual documents;
4) directories, catalogs and similar reference lists and tables; and
5) computer programs not requiring special characters.

Some of the requirements of text communication equipment conforming to this standard are that the equipment must be capable of: 1) receiving and storing all graphic characters of the repertoire without losing any of their significance; 2) presenting all graphic characters exactly or in an approximate form using fall-back techniques; 3) receiving all control functions of the repertoire; and, 4) performing the control functions exactly or storing them for later action, possibly using fall-back techniques.

6.2.2 Part 2 -

Part 2 of ISO 6937 defines the comprehensive repertoire of Latin alphabetic and non-alphabetic graphic characters for use in text communication; specifies coded representation for the graphic characters; and, specifies rules for the definition and use of subsets of the comprehensive repertoire of graphic characters. In addition, Part 2 has 3 annexes which are not part of the standard, but are intended to help clarify parts of the standard. Annex A summarizes the diacritical marks, along with their coded representation, and specifies the combinations of diacritical marks and basic letters which are defined in this part of the standard; Annex B is a survey of the use of the Latin alphabet characters in various languages; and, Annex C contains recommendations for fall-back presentations of graphic characters.

The coded representation of the graphic characters are divided into two sets: a primary set and a supplementary set. The primary set is intended to be used as the G0 set and is a subset of the graphic character set of the IRV of ISO 646.

The supplementary set is intended to be used as either the G1 or G2 set and is invoked as described in the revision of ISO 2022; word processing symbols are included in this category. The supplementary set contains three types of elements:

1) diacritical marks which form accented letters when used in combination with letters of the basic Latin alphabet from the primary set;

2) special alphabetic characters which are not included in the basic Latin alphabet and are not composed of diacritcal marks and basic letters (e.g., the German small sharp s and the Norwegian small ae ligature);

3) non-alphabetic characters not included in the primary set (e.g., section symbol, dollar sign, yen symbol).

For the purpose of communicating with equipment capable of presenting text using a limited set of graphic characters at one time, such as one which has an interchangeable printing element, this standard specifies certain rules for defining subrepertoires of the comprehensive set. Those rules include:

- Subrepertoires must include the 26 unaccented small and capital letters, the 10 decimal digits, and a specific list of special symbols defined in this standard.
- A subrepertoire may include, and is limited to, any additional graphic characters defined in this standard.
- The number of graphic characters is limited only by the size of the comprehensive repertoire.
- Two or more characters of the comprehensive set cannot be listed as a single character, nor can a single character of the comprehensive set having two names be assigned as two separate characters in the subrepertoire.

6.2.3 Part 3 -

Part 3 of ISO 6937 defines the comprehensive repertoire of control functions for text communication in page-image format; specifies coded representations for those control functions; and specifies code extension facilities to be used for the designation and invocation of graphic characters. The repertoire of control functions is divided into four groups:

1) Format effectors - control functions which cause the active position to move within the text area of a page and from page to page;

2) Presentation control functions - control functions which influence the appearance of the text (e.g., page format or character spacing);

3) Code extension control functions;

4) Miscellaneous control functions - control functions which do not fit in any of the preceding categories (e.g., Substitute character).

There are three types of control function sets: a primary set, a supplementary set, and an independent set. The elements of the primary set are represented in columns 0 and 1 in both 7- and 8-bit codes. The elements of the supplementary set are represented by escape sequences of the form ESC Fe in a 7-bit code and as single bit combinations in columns 8 and 9 in an 8-bit code. The elements of the independent set are represented by escape sequences of the form ESC Fs in both 7-bit and 8-bit codes.

The control functions included in the primary set are: Backspace, Line Feed, Form Feed, Carriage Return, Shift Out (Locking Shift 1 in the 8-bit code), Shift In (Locking Shift 0 in the 8-bit code), Substitute Character, and Escape. The control functions included in the supplementary set are: Partial Line Down, Partial Line Up, Reverse Line Feed, Single-Shift 2, Single-Shift 3, Page Separator, Document Separator, and Control Sequence Introducer. In addition, there are seven parameterized control function sequences defined in this standard. They are: Horizontal Position Relative, Vertical Position Relative, Select Graphic Rendition, Page Format Selection, Select Horizontal Spacing, Select Vertical Spacing, and Identify Graphic Subrepertoire. The formats of these control function sequences are identical to those described in section 5.4.

Several of the control function sequences are of particular interest to the office systems environment. They are:

1) Select Graphic Rendition (SGR) - a presentation control function with a selective parameter which specifies the appearance of the graphic characters in the subsequent text. Currently, the choices consist of a default, bold,

italics, underlined, or crossed-out (marked for deletion).

2) Page Format Selection (PFS) - a presentation control function with a selective parameter which specifies the format of subsequent pages. Some of the current choices are vertical 8-1/2 X 11, horizontal 8-1/2 X 11, vertical A4 (European size), and horizontal A4. This control function provides the implicit specification of margin stops.

3) Select Horizontal Spacing (SHS) - a presentation control function with a selective parameter which specifies the character spacing for subsequent text. Current choices include 10, 12, or 15 characters per 25.4mm.

4) Select Vertical Spacing (SVS) - a presentation control function with a selective parameter which specifies the line spacing for subsequent text. Current choices include 3, 4, 6, 8, or 12 lines per 25.4mm and 3, 4, 6, or 12 lines per 30.0mm.

5) Identify Graphic Subrepertoire (IGS) - a control function with one numeric parameter which is used to indicate to the receiving terminal that a particular subrepertoire is to be used in the subsequent text. This control function may be used to indicate to the receiver when to change the printing element on the output device.

The control functions included in the independent set are Locking Shift 1 Right, Locking Shift 2, Locking Shift 2 Right, Locking Shift 3, and Locking Shift 3 Right. Final character positions for these control functions have not yet been allocated.

6.3 CCITT Draft Recommendation S.61 (Character Repertoire and Coded Character Sets for the International Teletex Service)

Teletex is the name given to the internationally standardized text communications service between terminals which are used for the preparation, editing, and printing of correspondence (i.e., word processors and sophisticated electronic typewriters). The major objective of Teletex is to supply rapid convenient transmission of text with a low error rate and at a cost comparable to conventional mail. Teletex is expected to evolve into a service carrying both facsimile and textual information and will, therefore, play an important role in office automation [5].

Recommendation S.61 [15] is one of a set of Recommendations for the Teletex service which should be read in conjunction with each other. The other Recommendations are:

S.60 - Terminal equipment for use in the Teletex service;

S.62 - Control procedures for the Teletex service; and,
F.200 - Teletex service (fixes the rules to be followed).

Recommendation S.61 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. The development of the coded character set defined in S.61 is technically identical to ISO 6937, except that S.61 uses only an 8-bit structure for its code set. However, the following exceptions prevent the repertoires of graphic characters of S.61 and ISO 6937 from being completely compatible:

1) The number symbol and the international currency symbol are coded in positions 10/6 and 10/8, respectively, in the supplementary set of S.61, and are coded in positions 2/3 and 2/4, respectively, in the primary set of DP 6937; and,

2) S.61 provides for a non-spacing underline character (coded in position 12/12 of the supplementary set) for use with any other character in the graphic character repertoire, while DP 6937 does not.

The Teletex basic repertoire of control functions defined in S.61 consists of format effectors, presentation control functions, code extension control functions, and miscellaneous control functions. The format effectors are Space, Backspace, Line Feed, Form Feed, Carriage Return, Partial Line Up, and Partial Line Down. The presentation control functions are Page Format Selection, Select Graphic Rendition, Select Horizontal Spacing, and Select Vertical Spacing. The code extension control functions are Escape and Control Sequence Introducer. And, the miscellaneous control functions are Substitute Character and Identify Graphic Subrepertoire.

In addition to the basic level of Teletex, Recommendation S.61 also defines a higher level of service which consists of the basic level plus standardized options. It is intended that ultimately these standardized options will provide means for:

1) different character spacing;
2) different metric values for line spacing;
3) selection of different graphic renditions of text;
4) indication that special stationery should be used;
5) use of character repertoires other than the Teletex basic character repertoire (both national and application-oriented);
6) use of mixed modes of operation, such as facsimile coded information;
7) specification of an increased printable area; and,
8) escape into national and private options.

Currently, only part of these provisions have been defined through the addition of a few options to the basic repertoire of control functions (i.e., one format effector and additional parameter values for three of the presentation control functions). The remaining provisions are undergoing further study.

The optional format effector is Reverse Line Feed. Its intended use is to provide the capability of printing in the top few lines that are skipped by Form Feed.

The three presentation control functions which provide standardized options are Page Format Selection (PFS), Select Horizontal Spacing (SHS), and Select Vertical Spacing (SVS). This means that for PFS, parameter values are defined for the vertical and horizontal orientations of both a basic page format and the A4 page format. For SHS, a choice can be made between 10, 12, or 15 characters per 25.4mm by adding the standardized options. And, for SVS, the addition of standardized options provides parameter values for 3, 4, 6, 8, and 12 lines per inch and 3, 4, 6, and 12 lines per 30.0 mm.

The committee within CCITT responsible for the development of the Recommendations for Teletex service is Study Group VIII (Teletex Terminal Rapporteur's Group). Recommendations S.60, S.61, S.62, and F.200 were approved as an international standard at the 1980 meeting of the Plenary Assembly of the CCITT.

6.4 Draft ANS for Text Information Interchange in Page Image Format

This standard specifies the format effectors, presentation control functions, and other character-oriented control functions which affect the presentation of text in page image format [4]. Page Image Format (PIF) is defined as that basic method used in transmitting character coded text from one device to another such that no text processing is required at the receiver; the presentation then is in the format determined by the sender.

No graphic characters are defined in this standard. However, two levels of control functions are defined; i.e., level 1, and level 2. Level 1 is a subset of level 2, and level 2 is a subset of the CCITT Teletex service.

PIF level 1 assumes no intelligence and no mode changes. It is intended to provide for compatibility with all existing equipment, including "dumb" terminals and printers. The format effectors defined for level 1 are Space, Carriage Return, Line Feed, and Form Feed; one miscellaneous control character defined for level 1 is Substitute character. Page Image Format level 1 uses only a single page format which is based on the common printable area of both the ISO A4 and 8-1/2x11 inch paper sizes in the vertical paper orientation.

PIF level 2 is intended to be compatible with the basic level of Teletex. The format effectors defined for level 2 include those defined for level 1, plus Backspace, Partial Line Up and Partial Line Down. PIF level 2 also allows four presentation control functions (i.e., Page Format Selection, Select Graphic Rendition, Select Horizontal Spacing, and Select Vertical Spacing) and two code extension functions (i.e., Escape and Control Sequence Introducer). The page formats provided for in level 2 include the common printable areas of both the vertical orientation and the horizontal orientation of A4 and 8-1/2x11 inch paper sizes.

The committee within ANSI responsible for developing this standard is X3V1/WG4. It is the intention of the Committee that this standard will be one in a family of standards which will be compatible with the various levels of the CCITT Recommendations for the Teletex service. Therefore, this draft standard will be revised wherever necessary to become completely compatible with the CCITT Recommendation S.61.

7.0 SUMMARY

This report has described the status of the standardization efforts in coded character sets, especially relative to text processing, and the methods of extending those coded character sets. Also, it has demonstrated the effort of various national and international groups to develop standards which build on existing standards and which are compatible with the products of their standards-setting counterparts.

For the most part compatibility across standard organizations is maintained. However, at times the needs of the participants of different organizations vary. For example, some graphic characters in the International Reference Version code table did not meet certain communications needs for the United States; therefore, ANSI developed the ASCII code table to facilitate communications within the U.S. In the same vein, the CCITT in developing its Recommendations for Teletex Service assigned several graphic characters to positions not identical to the positions of corresponding characters in existing standards. Such changes result in incompatibility with the existing standards. Thus, the developers of subsequent standards must decide whether to remain compatible with other existing standards or to be compatible with the new CCITT Recommendations for Teletex. Due to the fact that the Teletex service will be available internationally, the trend appears to be to achieve compatibility with Teletex.

The standards discussed in this report accommodate a wide variety of functions required in text processing applications. However, certain limitations still exist when interchanging information via communicating text processors. These limitations are found in both the formatting and editing areas. The remainder of this report discusses these limitations and proposed solutions.

One of the most obvious omissions is the lack of controls to set physical margin stops. ISO DP 6937 implicitly sets margin stops according to page format (e.g., at positions 5 and 77 for vertical A4 format). Therefore, the addition of a control function in ISO DIS 6429 to explicitly set margin stops to conform to the assumed margin stops in DP 6937 would be a feasible solution to this omission [18].

Most text processing systems use hyphens in the text stream in two different ways; one is optional (also called a "soft" hyphen), and the other is a required hyphen. Optional hyphens are often used at the end of the line when a word will not fit; however, if, during reformatting, the word no longer falls at the end of the line, the hyphen will be automatically eliminated from the text. Required hyphens are generally input in combination with a special code key, thus making it a permanent part of the word (e.g., X-ray) and preventing the word from being broken across a line boundary. The existing standards do not provide the capability of inputting these two implementations of hyphen.

Similarly, many text processors distinguish between a space and a required space. A required space is used where the operator does not want a particular phrase broken across a line boundary (e.g., December 10). The required space is not defined in the existing standards.

Another major deficiency from the point of view of communicating text processors is the handling of text or control information outside the margins. Examples of information outside the margins are:

1) header and footer information;
2) when and where to start and stop page numbers (i.e., not on the title page);
3) notes inserted within the text to help the operator or author (the notes can be printed only at the request of the user); and,
4) footnote material meant to be printed on the bottom of the page on which the material is referenced.

It is possible that these applications can be provided through the use of APC (Applications Program Command) or OSC (Operating Systems Command); however, the standards are very vague as to the intended use of these control strings.

An application which is often used as a cost/benefit justification for text processing systems is the ability to increase productivity with the production of form letters. Form letters are commonly generated by merging variable information from a parameter file into a text template which also contains the control codes that control the merging process. Although it would not be reasonable to standardize the variable information file, it would be reasonable to standardize the control codes within the template file. This would allow organizations operating in various locations to communicate these template files between offices.

In a manner similar to that described for generation of form letters, most sophisticated text processors permit a user to key in text in an abbreviated form, touch a code key, and see the expanded version on their display. (The abbreviation and its expanded form have been predefined and stored in a separate file.) An extension to assist document interchange would be to have some mode switch which would store the abbreviation in the text (printed only at the request of the user) as well as the expanded text. Thus, when the document is transmitted, it could be displayed as it was actually input without transmitting the entire abbreviation glossary.

Existing communicating text processors are generally able to retain both the text and the control codes in the data stream when communicating documents to a system manufactured by the same vendor. Then, the receiving station can treat the document as if it were created there. However, the effect on information transmitted between different vendors' equipment varies widely. Most vendors try to preserve the graphic characters and, if possible, the original format. They often do this by transmitting an image of the document rather than the control codes themselves (e.g., changing a tab to 8 spaces and then transmitting the spaces rather than the tab). The document can then be printed out as intended by the sender; but, it is difficult or impossible to re-edit the document and transmit it back to the originator or another recipient.

This particular problem is currently being introduced into the standards arena as a new item of work. It is called Processable Text Form (PTF) and addresses the issue of retaining the control codes along with the graphic characters in information interchange, thereby allowing the receiver to process the document and retransmit it. The committees within ANSI charged with development of PTF standards are X3V1 and X3J6.

Certain other minor deficiencies in the present standards should be mentioned. They are:

1) A control function exists that allows a character to be repeated n times horizontally across a page. No control function exists, however, which allows vertical repetition. Clearly, this ability is useful when drawing forms.

2) The occurrence of PLU/PLD (Partial Line Up/Partial Line Down) is defined in non-nested pairs only. Many text processors are capable of multiple levels of super and subscripting. Therefore, the action of nested pairs of PLU/PLD should be well-defined in the standards.

3) No control exists to handle the decimal tab function which is present on all major vendors' word processing equipment. A decimal tab permits data presented in columns to be aligned along the decimal point.

Some of the deficiencies identified here can be alleviated in the standards addressing additional control functions (i.e., ISO DIS 6429, ANSI X3.64, and ECMA-48) by extending existing control functions and using the positions reserved for future use. The other deficiencies will be addressed during the development of the Processable Text Form standards. Attention to PTF will increase over the next few years as current draft standards near final approval.

REFERENCES

1. American National Standards Institute, *Additional Controls for Use with American National Standard Code for Information Interchange*, ANSI X3.64, 1979 July 18.

2. American National Standards Institute, *Code Extension Techniques for Use with the 7-Bit Coded Character Set of American National Standard Code for Information Interchange*, ANSI X3.41, 1974 May 14.

3. American National Standards Institute, *Code for Information Interchange*, ANSI X3.4, 1977 June 9.

4. American National Standards Institute, *Draft Proposed American National Standard for Text Information Interchange in Page Image Format (PIF)*, X3V1/81-19, 1981 February 2.

5. *The Future of Text Communications: The Impact of Teletex*. Mackintosh Publications Limited, Mackintosh House, Napier Road, Luton, LU1 1Rg, England. January 1980.

6. European Computer Manufacturers Association, *Additional Control Functions for Character-Imaging I/O Devices*, Standard ECMA-48, 2nd Edition, January 1979.

7. European Computer Manufacturers Association, *Extension of the 7-Bit Coded Character Set*, Standard ECMA-35, December 1971.

8. European Computer Manufacturers Association, *7-Bit Input/Output Coded Character Set*, Standard ECMA-6, 4th Edition, August 1973.

9. International Organization for Standardization, *Additional Control Functions for Character-Imaging Devices*, ISO DIS 6429, June 1978.

10. International Organization for Standardization, *Code Extension Techniques for Use with the ISO 7-Bit Coded Character Set*, ISO 2022, 1 July 1973.

11. International Organization for Standardization, *Code Extension Techniques for Use with the ISO 7-Bit Coded Character Set*, Revision of ISO 2022 (First ISO Draft Proposal), 2 August 1978.

12. International Organization for Standardization, *Coded Character Sets for Text Communication*, (Eighth Working Draft for an International Standard), ISO/TC97/SC2 N1074 (Part 1: General Introduction), N1075 (Part 2: Latin Alphabetic and Non-Alphabetic Graphic Characters), 9 January 1981; N1103 (Part 3: Control Functions for Page-Image Format), 31 March 1981.

13. International Organization for Standardization, Data Processing -- Procedure for Registration of Escape Sequences, ISO 2375, 1 July 1974.

14. Internantional Organization for Standardization, 7-Bit Coded Character Set for Information Processing Interchange, ISO-646, 1 July 1973.

15. International Telegraph and Telephone Consultative Committee, Character Repertoire and Coded Character Sets for the International Teletex Service, CCITT Draft Recommendation S.61 (1980).

16. International Telegraph and Telephone Consultative Committee, International Alphabet No. 5, CCITT Recommendation V.3 (1972).

17. International Telegraph and Telephone Consultative Committee, "Temporary Document No. 38," Source: Ad hoc group on S.f problems, Montreal, 2-6 June 1980.

18. National Bureau of Standards, An interim report done under contract by Advanced Office Concepts Corporation, 30 August 1980.

19. National Bureau of Standards, Code Extension Techniques in 7 or 8 Bits, Federal Information Processing Standards Publication 35, 1 June 1975.

20. National Bureau of Standards, Code for Information Interchange, Federal Information Processing Standards Publication 1-1, 24 December 1980.

21. National Bureau of Standards, Federal Information Processing Standards Index, Federal Information Processing Standards Publication 12-2, 1 December 1974.

22. Prigge, R.D., Marjorie F. Hill and Josephine L. Walkowicz, The World of EDP Standards, Sperry Univac, November 1978.

NBS-114A (REV. 2-80)

U.S. DEPT. OF COMM. BIBLIOGRAPHIC DATA SHEET (See instructions)	1. PUBLICATION OR REPORT NO. NBS SP 500-81	2. Performing Organ. Report No.	3. Publication Date September 1981

4. TITLE AND SUBTITLE

A Survey of Standardization Efforts of Coded Character Sets for Text Processing

5. AUTHOR(S)

Joan E. Knoerdel

6. PERFORMING ORGANIZATION (*If joint or other than NBS, see instructions*)

NATIONAL BUREAU OF STANDARDS
DEPARTMENT OF COMMERCE
WASHINGTON, D.C. 20234

7. Contract/Grant No.

8. Type of Report & Period Covered

Final

9. SPONSORING ORGANIZATION NAME AND COMPLETE ADDRESS (*Street, City, State, ZIP*)

Same as item 6.

10. SUPPLEMENTARY NOTES

Library of Congress Catalog Card Number: 81-600108

□ Document describes a computer program; SF-185, FIPS Software Summary, is attached.

11. ABSTRACT (*A 200-word or less factual summary of most significant information. If document includes a significant bibliography or literature survey, mention it here*)

As the desire to interchange documents among different text processing systems via communications increases, the incompatibilities which exist between those text processing systems become more and more apparent. One such incompatibility is that often the sending device and the receiving device use different bit assignments or coding schemes for their alphanumeric characters, special symbols, and control characters (i.e., different coded character sets). Considerable work has been done both nationally and internationally to standardize coded character sets. However, the knowledge of such standards efforts is not always widespread. Or, if there is familiarity with the standards efforts, the relationship of those efforts among various standards organizations is frequently not easy to perceive. The objective of this report is to describe the status of those standard coded character sets, with special attention to text processing systems. The report includes, first of all, a brief description of the major national and international standards organizations which develop code standards. Next, it describes the various code standards according to the following categories: basic code sets for information interchange, methods of augmenting those basic code sets, additional control characters to be used with the basic code sets, and code sets developed specifically for text communications. Finally, the summary of the report discusses a number of limitations which still exist when interchanging information via communicating text processors.

12. KEY WORDS (*Six to twelve entries; alphabetical order; capitalize only proper names; and separate key words by semicolons*)

Code extension techniques; code standards; coded character sets; control functions; information interchange; text communications; text processor.

13. AVAILABILITY

☒ Unlimited

□ For Official Distribution. Do Not Release to NTIS

☒ Order From Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402.

□ Order From National Technical Information Service (NTIS), Springfield, VA. 22161

14. NO. OF PRINTED PAGES

50

15. Price

$3.00

USCOMM-DC 6043-P80

☆ U.S. GOVERNMENT PRINTING OFFICE : 1981 O—340-997 (1784)

ANNOUNCEMENT OF NEW PUBLICATIONS ON COMPUTER SCIENCE & TECHNOLOGY

Superintendent of Documents,
Government Printing Office,
Washington, D. C. 20402

Dear Sir:

Please add my name to the announcement list of new publications to be issued in the series: National Bureau of Standards Special Publication 500-.

Name ______________________________

Company ______________________________

Address ______________________________

City ______________ State __________ Zip Code __________

(Notification key N-503)

NBS TECHNICAL PUBLICATIONS

PERIODICALS

JOURNAL OF RESEARCH—The Journal of Research of the National Bureau of Standards reports NBS research and development in those disciplines of the physical and engineering sciences in which the Bureau is active. These include physics, chemistry, engineering, mathematics, and computer sciences. Papers cover a broad range of subjects, with major emphasis on measurement methodology and the basic technology underlying standardization. Also included from time to time are survey articles on topics closely related to the Bureau's technical and scientific programs. As a special service to subscribers each issue contains complete citations to all recent Bureau publications in both NBS and non-NBS media. Issued six times a year. Annual subscription: domestic $13; foreign $16.25. Single copy, $3 domestic; $3.75 foreign.

NOTE: The Journal was formerly published in two sections: Section A "Physics and Chemistry" and Section B "Mathematical Sciences."

DIMENSIONS/NBS—This monthly magazine is published to inform scientists, engineers, business and industry leaders, teachers, students, and consumers of the latest advances in science and technology, with primary emphasis on work at NBS. The magazine highlights and reviews such issues as energy research, fire protection, building technology, metric conversion, pollution abatement, health and safety, and consumer product performance. In addition, it reports the results of Bureau programs in measurement standards and techniques, properties of matter and materials, engineering standards and services, instrumentation, and automatic data processing. Annual subscription: domestic $11; foreign $13.75.

NONPERIODICALS

Monographs—Major contributions to the technical literature on various subjects related to the Bureau's scientific and technical activities.

Handbooks—Recommended codes of engineering and industrial practice (including safety codes) developed in cooperation with interested industries, professional organizations, and regulatory bodies.

Special Publications—Include proceedings of conferences sponsored by NBS, NBS annual reports, and other special publications appropriate to this grouping such as wall charts, pocket cards, and bibliographies.

Applied Mathematics Series—Mathematical tables, manuals, and studies of special interest to physicists, engineers, chemists, biologists, mathematicians, computer programmers, and others engaged in scientific and technical work.

National Standard Reference Data Series—Provides quantitative data on the physical and chemical properties of materials, compiled from the world's literature and critically evaluated. Developed under a worldwide program coordinated by NBS under the authority of the National Standard Data Act (Public Law 90-396).

NOTE: The principal publication outlet for the foregoing data is the Journal of Physical and Chemical Reference Data (JPCRD) published quarterly for NBS by the American Chemical Society (ACS) and the American Institute of Physics (AIP). Subscriptions, reprints, and supplements available from ACS, 1155 Sixteenth St., NW, Washington, DC 20056.

Building Science Series—Disseminates technical information developed at the Bureau on building materials, components, systems, and whole structures. The series presents research results, test methods, and performance criteria related to the structural and environmental functions and the durability and safety characteristics of building elements and systems.

Technical Notes—Studies or reports which are complete in themselves but restrictive in their treatment of a subject. Analogous to monographs but not so comprehensive in scope or definitive in treatment of the subject area. Often serve as a vehicle for final reports of work performed at NBS under the sponsorship of other government agencies.

Voluntary Product Standards—Developed under procedures published by the Department of Commerce in Part 10, Title 15, of the Code of Federal Regulations. The standards establish nationally recognized requirements for products, and provide all concerned interests with a basis for common understanding of the characteristics of the products. NBS administers this program as a supplement to the activities of the private sector standardizing organizations.

Consumer Information Series—Practical information, based on NBS research and experience, covering areas of interest to the consumer. Easily understandable language and illustrations provide useful background knowledge for shopping in today's technological marketplace.

Order the **above** *NBS publications from: Superintendent of Documents, Government Printing Office, Washington, DC 20402.*

Order the **following** *NBS publications—FIPS and NBSIR's—from the National Technical Information Services, Springfield, VA 22161.*

Federal Information Processing Standards Publications (**FIPS PUB**)—Publications in this series collectively constitute the Federal Information Processing Standards Register. The Register serves as the official source of information in the Federal Government regarding standards issued by NBS pursuant to the Federal Property and Administrative Services Act of 1949 as amended, Public Law 89-306 (79 Stat. 1127), and as implemented by Executive Order 11717 (38 FR 12315, dated May 11, 1973) and Part 6 of Title 15 CFR (Code of Federal Regulations).

NBS Interagency Reports (**NBSIR**)—A special series of interim or final reports on work performed by NBS for outside sponsors (both government and non-government). In general, initial distribution is handled by the sponsor; public distribution is by the National Technical Information Services, Springfield, VA 22161, in paper copy or microfiche form.