DOCUMENT RESUME

ED 088 485 IR 000 319

AUTHOR	Duncan, Blanton C.; Garvin, David
TITLE	Complete Clear Text Representation of Scientific Documents in Machine-Readable Form. NBS Technical Note 820.
INSTITUTION	National Bureau of Standards (DOC), Washington, D.C.
REPORT NO	NBS-TN-820
PUB DATE	Feb 74
NOTE	56p.
AVAILABLE FROM	Superintendent of Documents, U.S. Government Printing Office, Washington, D. C. 20402 (SD Catalog No. C13.46:820; $.90)
EDRS PRICE	MF-$0.75 HC-$3.15
DESCRIPTORS	Communications; *Computer Graphics; *Computer Storage Devices; Data Analysis; Data Processing; Information Centers; Information Dissemination; Information Processing; *Information Science; *Machine Translation; Program Descriptions; Sciences; *Symbols (Mathematics); Technology
IDENTIFIERS	Clear Text Representation; Graphic Character Sets; Information Interchange Codes; International Organization for Standardization; ISO; National Bureau of Standards; NBS

ABSTRACT

Science and technology use a large variety of symbols to represent physical properties, chemical formulas, and mathematical expressions. Since data centers which codify and evaluate physical properties need to use this conventional symbolism, it is recommended that they adopt the symbols and terminology specified by the various International Unions, both for manual operations and for the creation of machine-readable data bases. It is demonstrated that these conventional symbols can be produced by communications devices which are compatible with the international standard codes for information exchange. A set of characters suitable for representing scientific data and text is presented and proposed as an extension of the International Organization for Standardization(ISO) information interchange code. The use of this extended character code by computer oriented data centers at the National Bureau of Standards is described. In addition, the kinds of equipment needed for this level of performance and criteria for their selection are outlined. (Author)

A UNITED STATES DEPARTMENT OF COMMERCE PUBLICATION

NBS TECHNICAL NOTE 820

Complete Clear Text Representation of Scientific Documents in Machine-Readable Form

U.S. DEPARTMENT OF COMMERCE
National Bureau of Standards

NATIONAL BUREAU OF STANDARDS

The National Bureau of Standards[1] was established by an act of Congress 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 consists of the Institute for Basic Standards, the Institute for Materials Research, the Institute for Applied Technology, the Institute for Computer Sciences and Technology, and the Office for Information Programs.

THE INSTITUTE FOR BASIC STANDARDS provides the central basis within the United States of a complete and consistent system of physical measurement; coordinates that system with measurement systems of other nations; and furnishes essential services leading to accurate and uniform physical measurements throughout the Nation's scientific community, industry, and commerce. The Institute consists of a Center for Radiation Research, an Office of Measurement Services and the following divisions:

Applied Mathematics — Electricity — Mechanics — Heat — Optical Physics — Nuclear Sciences[2] — Applied Radiation[2] — Quantum Electronics[3] — Electromagnetics[3] — Time and Frequency[3] — Laboratory Astrophysics[3] — Cryogenics[3].

THE INSTITUTE FOR MATERIALS RESEARCH conducts materials research leading to improved methods of measurement, standards, and data on the properties of well-characterized materials needed by industry, commerce, educational institutions, and Government; provides advisory and research services to other Government agencies; and develops, produces, and distributes standard reference materials. The Institute consists of the Office of Standard Reference Materials and the following divisions:

Analytical Chemistry — Polymers — Metallurgy — Inorganic Materials — Reactor Radiation — Physical Chemistry.

THE INSTITUTE FOR APPLIED TECHNOLOGY provides technical services to promote the use of available technology and to facilitate technological innovation in industry and Government; cooperates with public and private organizations leading to the development of technological standards (including mandatory safety standards), codes and methods of test; and provides technical advice and services to Government agencies upon request. The Institute consists of a Center for Building Technology and the following divisions and offices:

Engineering and Product Standards — Weights and Measures — Invention and Innovation — Product Evaluation Technology — Electronic Technology — Technical Analysis — Measurement Engineering — Structures, Materials, and Life Safety[4] — Building Environment[4] — Technical Evaluation and Application[4] — Fire Technology.

THE INSTITUTE FOR COMPUTER SCIENCES AND TECHNOLOGY conducts research and provides technical services designed to aid Government agencies in improving cost effectiveness in the conduct of their programs through the selection, acquisition, and effective utilization of automatic data processing equipment; and serves as the principal focus within the executive branch for the development of Federal standards for automatic data processing equipment, techniques, and computer languages. The Institute consists of the following divisions:

Computer Services — Systems and Software — Computer Systems Engineering — Information Technology.

THE OFFICE FOR INFORMATION PROGRAMS promotes optimum dissemination and accessibility of scientific information generated within NBS and other agencies of the Federal Government; promotes the development of the National Standard Reference Data System and a system of information analysis centers dealing with the broader aspects of the National Measurement System; provides appropriate services to ensure that the NBS staff has optimum accessibility to the scientific information of the world. The Office consists of the following organizational units:

Office of Standard Reference Data — Office of Information Activities — Office of Technical Publications — Library — Office of International Relations.

---

[1] Headquarters and Laboratories at Gaithersburg, Maryland, unless otherwise noted; mailing address Washington, D.C. 20234.
[2] Part of the Center for Radiation Research.
[3] Located at Boulder, Colorado 80302.
[4] Part of the Center for Building Technology.

Complete Clear Text Representation of Scientific Documents in Machine-Readable Form

Blanton C. Duncan

Computer Services Division
Institute for Computer Sciences and Technology

and

David Garvin

Physical Chemistry Division
Institute for Materials Research

National Bureau of Standards
Washington, D.C. 20234

U.S. DEPARTMENT OF COMMERCE, Frederick B. Dent, *Secretary*

NATIONAL BUREAU OF STANDARDS, Richard W. Roberts, *Director*

Issued February 1974

National Bureau of Standards Technical Note 820

Nat. Bur. Stand. (U.S.), Tech. Note 820, 55 pages (Feb. 1974)

CODEN: NBTNAE

U.S. GOVERNMENT PRINTING OFFICE
WASHINGTON: 1974

For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402
(Order by SD Catalog No. C13.46:820). Price 90 cents.

Preface

Ordinary text is the most common class of non-numeric data to be manipulated, and scientific text and diagrams are among the most complex. This Technical Note is intended as a contribution to the development of standards which will permit the facile dissemination of all types of non-numeric data in machine-readable form.

Publication of this material as an NBS Technical Note has the advantage that the finished document itself stands as an illustration of the major points discussed. Except for the cover, title page and bibliographic record, the physical presentation is a copy of a "print-out" from magnetic and paper tapes of an edited manuscript plus figures in the form in which it would be seen by editorial and technical reviewers prior to publication. There are only two exceptions as to detail:

(1) The bracketing data contained in the machine record and on the margins of this page (line numbers, y coordinates, etc.) is not printed in the document proper which follows;

(2) Some of the arts of composition and make-up have been applied to the extent that figures and their captions have been merged into the document at the proper places rather than retained as separate collections of data elements to be manipulated by the make-up editors themselves.

In a sense, this entire NBS Technical Note is a single "figure" which might be included as part of a publication describing the data content of a reel of magnetic tape. At the present time we cannot anticipate any real use for a magnetic tape version of this Note.

On the other hand, authors of technical documents cannot be sure, today, that there will be no demand for a machine readable copy of

..5...10...15...20...25...30...35...40...45...50 Max y = 136

their work. Thus they should have facilities - 3 1
for encoding their documents in full detail - 6 2
against this eventual demand. We describe here - 9 3
a technique that permits this and has enabled - 12 4
us to meet such demands. - 15 5

Blanton C. Duncan ± 20 6
David Garvin - 23 7

Table of Contents

Figures

Complete clear text representation of scientific documents in machine-readable form

by
Blanton C. Duncan and David Garvin
National Bureau of Standards, (U.S.)
Washington, D. C.

Science and technology use a large variety of symbols to represent physical properties, chemical formulas and mathematical expressions. Data centers that codify and evaluate physical properties need to use this conventional symbolism in their work. It is recommended that these data centers adopt the symbols and terminology specified by the various International Unions both in manual operations and in the creation of machine-readable data bases.

It is demonstrated that these conventional symbols can be produced by modern communications devices that are compatible with the international standard codes for information interchange. A set of characters suitable for representing scientific data and text is presented and proposed as an extension of the ISO information interchange code.

The use of this extended character code by computer oriented data centers at the National Bureau of Standards is described. The equipment needed for this level of performance and criteria for their selection are outlined.

Key Words: graphic character sets; information analysis centers; information interchange codes; recording typewriters; scientific computer technology.

1. Introduction

The problem of codifying the results of scientific research has received increased emphasis in recent years. The task is immense. Data produced throughout the world must be assembled, analyzed by experts and the best possible results be made available to the ultimate user in a useful form. A promising approach which has received widespread attention is the establishment of a large number of data analysis centers, each devoted to a specialty. As the number of centers grows,

so will the need for them to trade information, often across national boundaries. The centers will also need to provide information to remote users and do this quickly. This means that a substantial communications problem arises among a large number of independent groups.

Coupled to these problems of data analysis and communications is that of automation of data centers. The use of computers for data collection, reduction and analysis is widespread. Data centers need, in addition, text processing and file handling techniques suitable for the material they must collect, index, store and analyze.

This paper treats a basic subject: the recording of scientific data and text in an automated environment. It describes a `man-machine' alphabet or symbolism for the interchange and processing of scientific data: the General Purpose Scientific Document Code. The needs of the human are met by providing a set of symbols suitable for producing the basic scientific document--the typescript of a paper. The machine needs are met by associating this `alphabet' with the existing international standard system for information interchange.

The orientation of this paper is toward human requirements. The machine is to serve man, not vice-versa. To this end, emphasis will be placed on the signs, or graphic characters, used in written communications. The control elements necessary for the machine manipulation of this man-machine alphabet will be treated lightly, except where it is essential that they be mentioned.

This work has been, and is experimental. The context in which it was done controlled many of the decisions that were made. This background is the interaction of several small, independently managed data analysis centers at the National Bureau of Standards (NBS) both among themselves and with a general service computer center. The data centers and the computer services center have distinctly different problems.

The data centers, on the basis of substantial experience operating in a non-automated mode, have specified their text handling requirements. These run far beyond the facilities for processing and printing commonly available in computer centers. One example is the requirement for both capital and small Latin letters for representing the chemical elements. Another is the need for both superscripts and subscripts in mathematical, chemical, physical property and spectroscopic notation. The needs of the data centers are consistent with, and can be tested against the extensive work of international scientific organizations to define symbols, terminology and

nomenclature. These standardized symbols are displayed and defined in ISO, IUPAP, and IUPAC documents published during the 1960's [1,2,3,4].

From the viewpoint of a general service computer center, these are highly specialized requirements of a minor group of its customers. By definition (or by default) the majority of its existing customers can accept constraints imposed by keypunches and printers with limited character sets. However, the growing needs of data centers must be met. The facilities they require can be expected to have wider application, but a computer center cannot accept the responsibility to support an absolutely open-ended man-machine alphabet. A finite solution is needed.

The systematic development of a suitable, finite coded character set became practical in April 1965 with the publication of a proposed revision of the American National Standard Code for Information Interchange (ASCII) [5]. ASCII is an anticipatory standard. Even today the hardware and software of many general service computer centers are not designed to handle flows of data between man and machine at the levels of complexity anticipated by basic ASCII. But that standard, and the international one to which it is closely related, provide a carefully defined system within which the needs of science can be met. The standards can be used by computer centers for the planning of improvements in their services.

Thus the standards developed for science and those developed for communications can be brought to bear on a solution. However, standards are not enough. They must be implemented in hardware and software. The crucial parts are input devices to record the data, processing programs to accept, edit, reformat, retrieve, and store the data, and output machines to print clear, complete scientific text.

An experimental processing system based on the concepts described in this paper has been in use, on a production basis, by the NBS data centers since 1967. This "field testing" and the on-going standardization activities in the communications field have forced changes in detail but not in concept.

Parts of this work have been described before. Others are summarized here for the first time. A prototype input-output device, the "taxywriter", was developed to demonstrate the feasibility of the system [6]. Although superseded by commercially available instruments, it remains in service. The first version of this scientific man-machine alphabet was described in 1968 [7]. Comments received and further study led to a revised set of graphics. These were incorporated in a line printer installed at NBS [8].

From the human point of view this printer is the most important development. It can display fully a scientific typescript. An allied development is the use of photo-composition for final output. Surprisingly enough, this system, although designed for the typescript, is sufficiently rich to provide the printed results expected by the scientist.

The sections that follow treat a variety of topics. The objective is an overview of the subject. Concepts and criteria are emphasized at the expense of detail. The symbolism of science is displayed (Section 2). An introduction to the communications codes is provided. How these may be extended to meet scientific needs is explained (section 3). A specific extension, the General Purpose Scientific Document Code, is described together with examples of its use (section 4). Criteria for selection of input devices and the design of printers are discussed in so far as they bear on the man-machine interface.

It is appropriate to state here the conlusions that we have drawn from this work, or, if you will, display our biases. These are:

(a) The currently used complex terminology and symbols of science and technology are needed to represent the wide range of properties that are measured and the theories that interpret them.

(b) Data centers will need to use this symbolism in their internal work and in communication with others. In the present state of the art, automation equipment and computer techniques present no insuperable barriers to the use of this complex symbolism.

(c) Techniques for automated handling of scientific text can be developed within the context of international standards for information interchange. Cooperative development is possible on this basis.

(d) General concepts and criteria for text handling can be developed for the design of equipment and implementation of operating systems.

(e) The human factor, not the machine, should control the development. Ease of operation and flexibility must have first priority.

(f) A system must be expandable to meet future demands. The present is always an approximation.

2. The Symbols of Science

With what type of material must one contend in scientific communication? The determination must take into account the formal recommendations in standards documents [1,4], publication practices, and manuals of style [9, 10, 11]. The "IUPAC Manual for Symbols and Terminology for Physicochemical Quantities" [4] displays the formal recommendations very well. The other standard documents enlarge the set slightly. Publication practices have been sampled by examining publications of the American Chemical Society, the Association for Computing Machinery, the American Institute of Physics, the American Association for the Advancement of Science and the National Bureau of Standards. Among the style manuals, the "Handbook for Authors" of the American Chemical Society is particularly important [11]. It devotes considerable attention to the preparation of a scientific typescript. It gives many illustrations of the thesis developed below.

2.1 Information Content of a scientific article. Each article in a technical journal reaches the editor and the technical reviewers in the form of a typescript. This typewritten copy had (or should have had) all the symbols, equations, chemical formulas, etc. in a form readily recognizable by the typesetter. This leads to a general rule:

> The copy that can be produced by a scientific typist contains all the information that appears in written scientific communications. The minimum acceptable level of performance of a text-handling system for science is full reproduction of a scientific typescript.

2.2 Classes of symbols needed. The IUPAC definitions for the display of physicochemical quantities show that the complete Latin and Greek alphabets both in capital and small letters are necessary. A variety of special signs are needed to denote operations, relations, etc. Special relationships among symbols (such as the use of superscripts and subscripts) are prescribed. These have high information content.

The IUPAC definitions also show that typescript notations are required to supply stylistic information for the typesetter in order to permit the following rules to be implemented.

> Upright (roman) type is used for chemical formulas (section 7.2, ref 4), units (section 3.1) and mathematical operators (section 6).
>
> Slanting (italic) type is used for symbols for physical properties. These symbols are letters of the Latin and Greek alphabets (section 1.5). Vector quantities are printed in heavy (bold-face) slanting type (section 1.5).
>
> Superscripts and subscripts which are themselves symbols for physical properties are printed in slanting type. All others are printed in upright type (section 1.6).

Typescript notations to meet the requirements of these rules are discussed in section 4.3 of this paper.

The scientific typist needs two other classes of symbols. The first is a set of line segments for ruling tables, representing structures of organic chemicals, producing flow charts, etc. The second is a miscellany of marks ordinarily used in preparing text, such as punctuation, currency symbols and diacritical marks.

An automated system can usefully provide an additional feature. This is a set of dots for plotting. These are more likely to be used with computer generated data than by a typist.

Figure 1 shows a set of symbols useful for science.* Figure 2 shows several diagrams subject to being keyboarded.

2.3 Scientific communication. It has been a difficult task to define symbols and terminology for science. This task has occupied committees of the various scientific unions for decades. These committees have been successful

---

*Figure 1 is a photographic reduction of original copy produced at 3:1 scale on a computer driven incremental plotter using output from a typewriter simulator program. This design tool is discussed further in Section 4.8.

SCIENTIFIC NOTATION

Latin and Greek Alphabets 86 symbols

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

a b c d e f g h i j k l m n o p q r s t u v w x y z

Δ Φ Γ Ξ Λ Π Θ Σ Υ Ω Ψ

α β δ ε φ γ η ι ξ κ λ μ ν π θ ρ σ τ υ ω χ ψ ζ

Arabic Numerals 10 symbols

1 2 3 4 5 6 7 8 9 0

Mathematical Operators and Miscellaneous ~60 symbols

= ± ≠ ≡ ≙ ≈ → ≃ ∝ < > ≤ ≥ + − × / ∣ √ ∫ ! * . , : ; ° ' ′

( ) ÷ [ ] ~ ^ ≉ ∂ ∇ ∥ ≒ £ ¤ ¤ ϖ ⊂ ⊃ ∩ ∪ ← ¶ § ↑ ↓ ∃ ¬ ◊ #

¯ – & $ [ ] " % `

Line Segments and Dots for Diagrams 20 symbols

∣ • • . • ⟍ ⟍ ⟋ ⟋ ∣ • / / \ \ ⟋ ⫽ \ ⩥ ∣

USAGE: Letters may be upright (roman) and sloping (italic) in light or heavy (bold) type. All symbols may be used as superscripts and subscripts. Scientific typescripts reflect this usage, e.g.

$^{14}N_2^+$ C_p $\underset{\sim}{E} = \underset{\sim}{\rho} \bullet \underline{j}$ $\Delta H f^\circ$ $\Delta(N\Theta_3^-)$

Figure 1: Illustration of the notation employed in scientific typescripts when the typist has an adequate set of auxiliary type characters for the typewriter.

200 mm

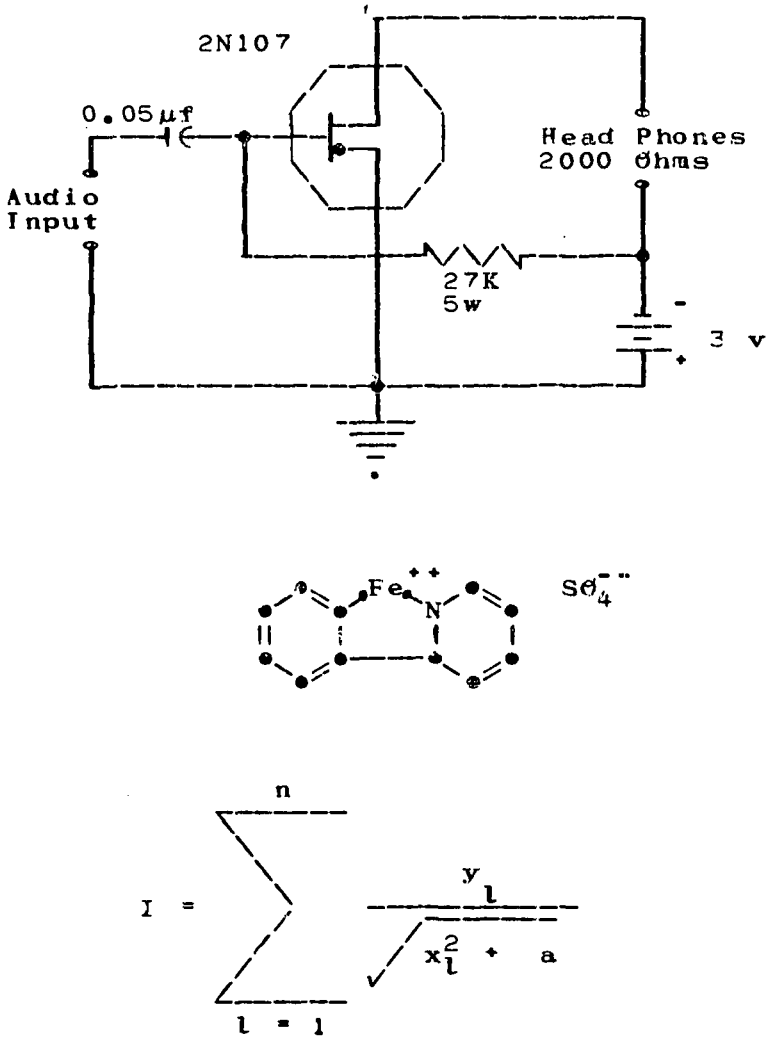
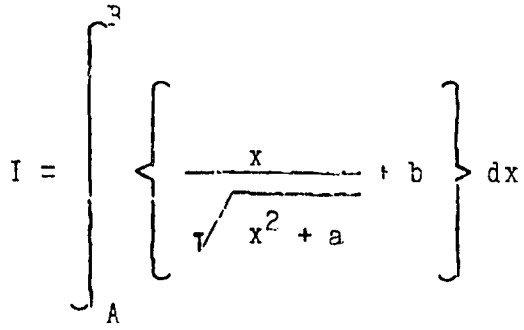
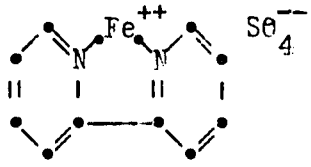


Figure 2

Diagrams. An electrical circuit diagram, chemical formulas, mathematical display equations, a graph and a block diagram. The first three (above) were produced as output on a computer driven line printer. The remainder (following two pages) were produced on a teletypewriter operating under punched paper tape control.





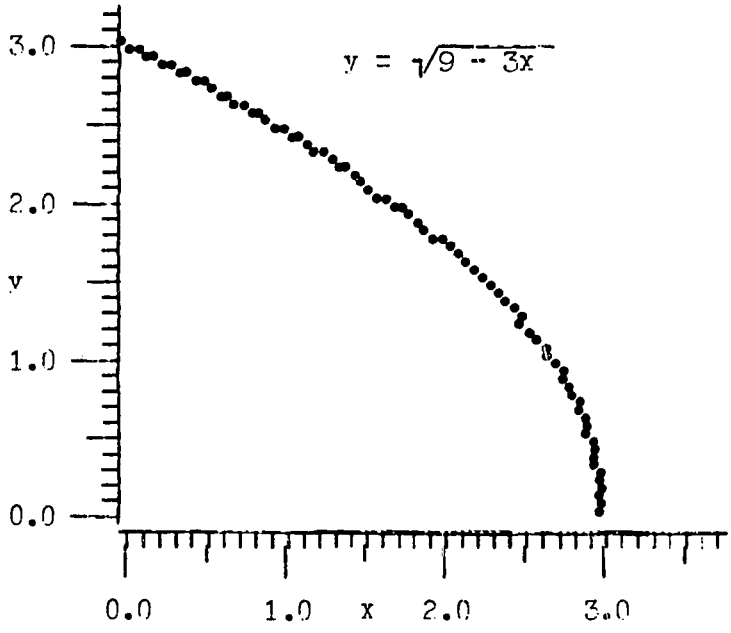
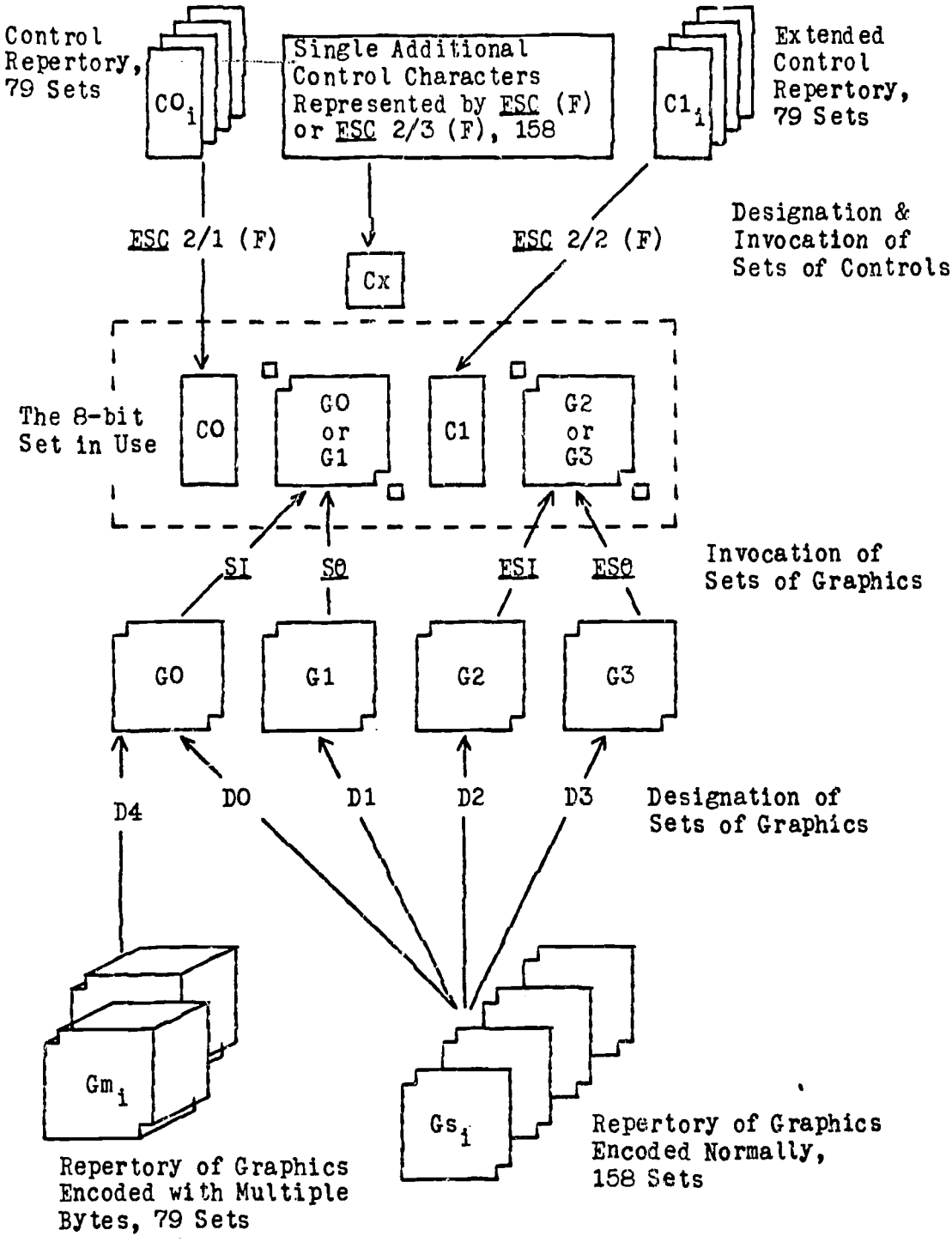


Figure 2. (continued)

STRUCTURE OF THE FAMILY OF 8-BIT CODES



Designation of Sets of Graphics:

DO = ESC 2/8 or 2/12 (F) D1 = ESC 2/9 or 2/13 (F)
D2 = ESC 2/10 or 2/14 (F) D3 = ESC 2/11 or 2/15 (F)
D4 = ESC 2/4 (F)

Figure 2. (concluded)

in providing a carefully defined comprehensive notation. It should be used.

Data centers should use this notation simply because it is their mission to codify the results of scientific investigation. It is reasonable to expect them to lead the way in standardizing the transfer of scientific information. Specialists in automation and information exchange must also consider these notational schemes. They are existing, not projected, procedures. They show demands that may be expected to increase in intensity in the future.

2.4 Summary. Science and technology transfer information using systems of notation that are far more complex than those required for newspapers, magazines and telegrams. On the other hand, all except the last of these common media for information interchange use symbolisms that are beyond the de facto standard output provided by common computer printers. The next section shows that the common computer facilities are sub-standard.

3. The ISO Code for Information Processing Interchange

The International Organization for Standardization Code for Information Processing Interchange, ISO R646, provides for and is already widely used in telecommunications [12, 13]. The American National Standard Code (ASCII-1968) is a proper variant [5, 14]. These codes and the doctrines for their use are still under development. Revisions are being considered, mainly to clarify matters of national use and the "national option" positions. But none of the proposed revisions would change the basic system upon which the work reported here rests.

Today, the ISO Code is a 7-bit code (128 patterns) with 33 control functions, "space" and 94 (visible) printing characters. The control functions provide for communication needs: "enquire", "acknowledge", "end of transmission", etc; represent typewriter operations: "line feed", "carriage return", "horizontal tabulation", "backspace", etc; and include a few information markers: file, group, record and unit separators. Each control function is carefully defined. We have found it possible to translate unambiguously into the standard those important control features of typewriters that employ other code schemes.

The 94 graphics characters of the ISØ code are a slight expansion of sets normally found on typewriters for the Latin alphabet. The set includes the capital and small Latin letters, numerals, punctuation marks, mathematical operators and a variety of special symbols.

The ISØ code is shown in the left hand side of Figure 3a starting with the character NUL and continuing through the character DEL in the columns labeled 0 through 7. It is suitable for simple text. Since the bulk of most scientific papers is simple text, the basic facilities of the ISØ code cannot be spared. It is not possible to replace any usefully large number of symbols of the basic set with ones more useful for scientific work.

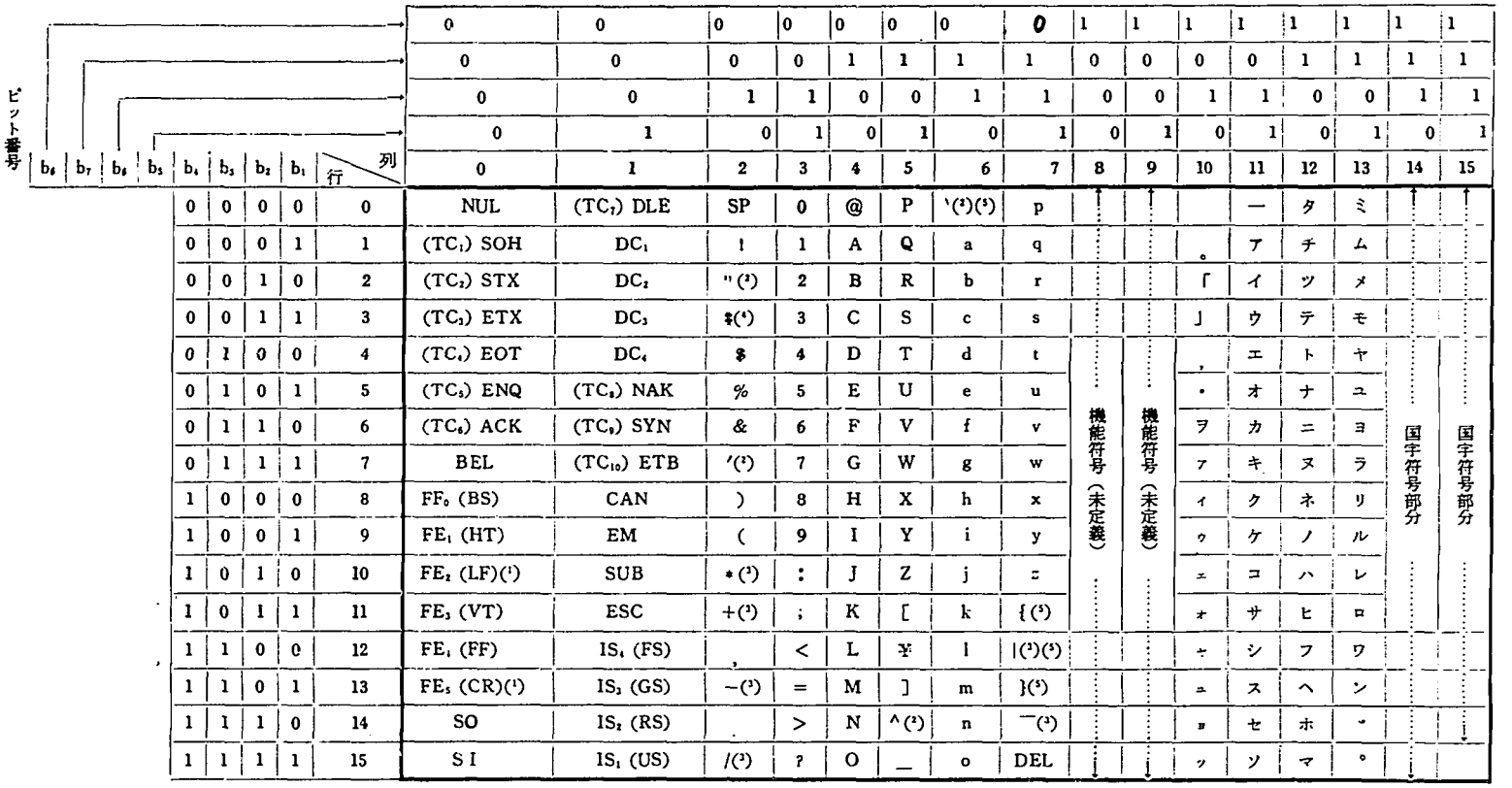
FIGURE 3

A family of codes related to the ISØ Information Processing Code. All are displayed in the same arrangement. Columns 0-7 have the basic 7-bit code, and columns 8-15 have extensions. These two regions are arranged in the same manner: columns 0 and 1 (8 and 9) are for controls, columns 2-7 (10-15) contain graphic symbols.

a. Japanese Industrial Standard Code [15]. The extension (columns 10-13) provides Katakana characters as an alternative set.

b. USSR Alpha-Numeric Code [16]. The extension repeats columns 2 and 3 in columns 10 and 11. Cyrillic characters, columns 12-15 are arranged to match their Latin equivalents in columns 4-7 where possible (but with capital Cyrillic overlaying small Latin letters).

c. American National Standard Code for Information Interchange (ASCII) and an extension for library use. Columns 2-7 give the basic ASCII set [14]. The extension provides a large collection of diacritical marks and some infreqntly used letters. This combined set is used as an 8-bit code by the U.S. Library of Congress [17].



ビット番号								
b_8	b_7	b_6	b_5	b_4	b_3	b_2	b_1	行＼列
0	0	0	0	0	0	0	0	0
0	0	0	0	1	1	1	1	1
0	0	0	1	0	0	1	1	2
0	0	0	1	1	0	1	0	3
0	0	1	0	0	1	0	1	4
0	0	1	0	1	1	0	0	5
0	0	1	1	0	1	1	0	6
0	0	1	1	1	1	1	1	7
0	1	0	0	0	0	0	0	8
0	1	0	0	1	0	0	1	9
0	1	0	1	0	0	1	0	10
0	1	0	1	1	0	1	1	11
0	1	1	0	0	1	0	0	12
0	1	1	0	1	1	0	1	13
0	1	1	1	0	1	1	0	14
0	1	1	1	1	1	1	1	15

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
NUL	(TC$_7$) DLE	SP	0	@	P	`(2)(5)	p				—	タ	ミ		
(TC$_1$) SOH	DC$_1$	!	1	A	Q	a	q			。	ア	チ	ム		
(TC$_2$) STX	DC$_2$	"(2)	2	B	R	b	r			「	イ	ツ	メ		
(TC$_3$) ETX	DC$_3$	#(4)	3	C	S	c	s			」	ウ	テ	モ		
(TC$_4$) EOT	DC$_4$	$	4	D	T	d	t			、	エ	ト	ヤ		
(TC$_5$) ENQ	(TC$_8$) NAK	%	5	E	U	e	u			・	オ	ナ	ユ		
(TC$_6$) ACK	(TC$_9$) SYN	&	6	F	V	f	v			ヲ	カ	ニ	ヨ		
BEL	(TC$_{10}$) ETB	'(1)	7	G	W	g	w			ァ	キ	ヌ	ラ		
FE$_0$ (BS)	CAN	)	8	H	X	h	x			ィ	ク	ネ	リ		
FE$_1$ (HT)	EM	(	9	I	Y	i	y			ゥ	ケ	ノ	ル		
FE$_2$ (LF)(1)	SUB	*(3)	:	J	Z	j	z			ェ	コ	ハ	レ		
FE$_3$ (VT)	ESC	+(3)	;	K	[	k	{(5)			ォ	サ	ヒ	ロ		
FE$_4$ (FF)	IS$_4$ (FS)	,	<	L	¥	l	|(5)			ャ	シ	フ	ワ		
FE$_5$ (CR)(1)	IS$_3$ (GS)	−(3)	=	M	]	m	}(5)			ュ	ス	ヘ	ン		
SO	IS$_2$ (RS)	.	>	N	^(3)	n	¯(3)			ョ	セ	ホ	゛		
SI	IS$_1$ (US)	/(3)	?	O	_	o	DEL			ッ	ソ	マ	゜		

機能符号（未定義）

機能符号（未定義）

国字符号部分

国字符号部分

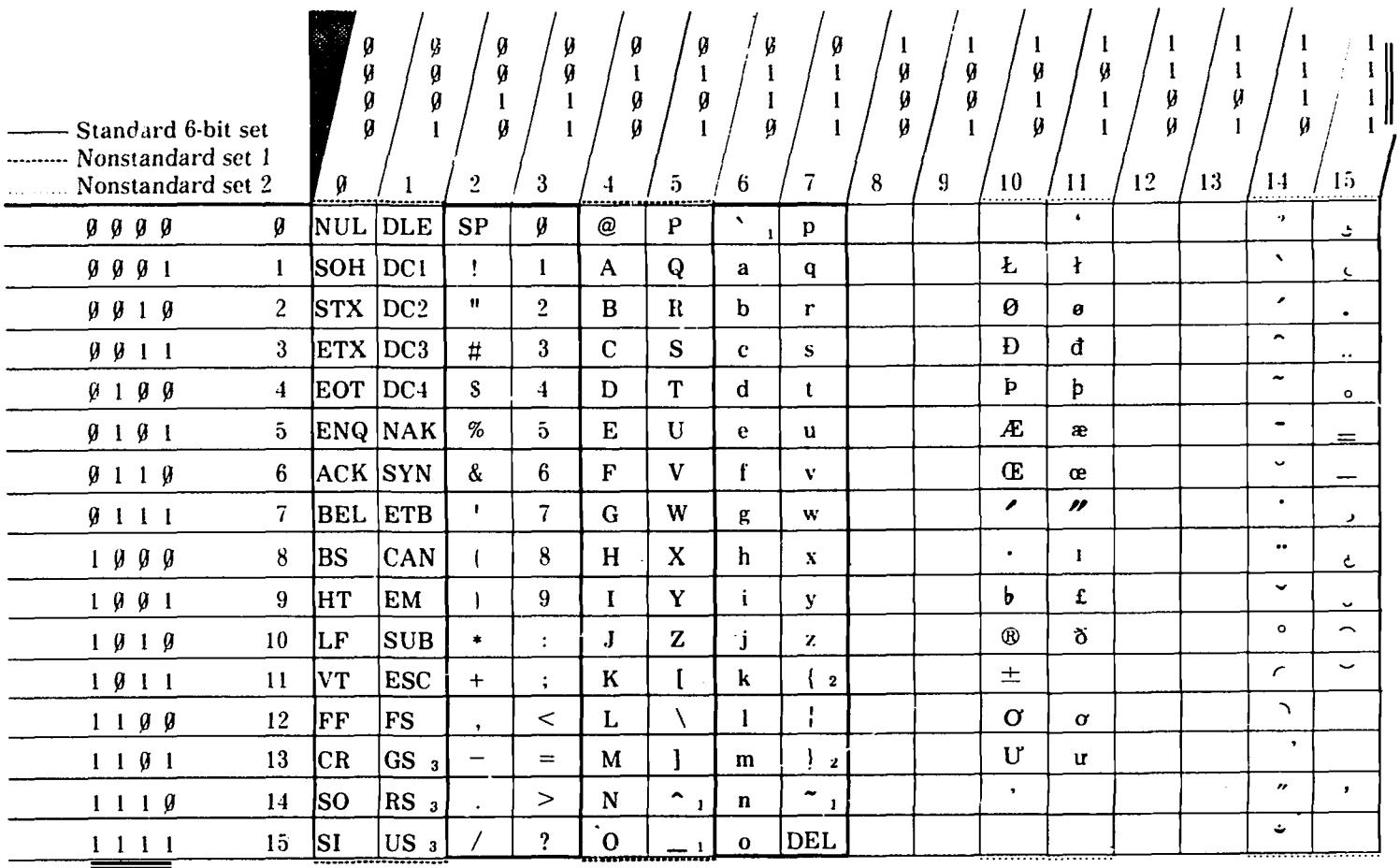
注 (1)(2)(3)(4)(5) 表1の注に同じ

Figure 3a. ISO Code with Japanese National options and extensions. The use of the Dollar Sign at column 2, row 4 and of the Yen Sign at column 5, row 12 are the only substitutions made with respect to the International Reference Version of the ISO Code.

		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
		0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	0	NUL	DLE	SP	0		P	`	p				0	ю	п	Ю	П
1	1	SOH	DC1	!	1	A	Q	a	q			!	1	а	я	А	Я
2	2	STX	DC2	"	2	B	R	b	r			"	2	б	р	Б	Р
3	3	ETX	DC3	#	3	C	S	c	s			#	3	ц	с	Ц	С
4	4	EOT	DC4	¤	4	D	T	d	t			¤	4	д	т	Д	Т
5	5	ENQ	NAK	%	5	E	U	e	u			%	5	е	у	Е	У
6	6	ACK	SYN	&	6	F	V	f	v			&	6	ф	ж	Ф	Ж
7	7	BEL	ETB	'	7	G	W	g	w			'	7	г	в	Г	В
8	8	BS	CAN	(	8	H	X	h	x			(	8	х	ь	Х	Ь
9	9	HT	EM	)	9	I	Y	i	y			)	9	и	ы	И	Ы
10	A	LF	SUB	*	:	J	Z	j	z			*	:	й	з	Й	З
11	B	VT	ESC	+	;	K	[	k				+	;	к	ш	К	Ш
12	C	FF	IS4	,	<	L	~	l				,	<	л	э	Л	Э
13	D	CR	IS3	-	=	M	]	m				-	=	м	щ	М	Щ
14	E	RUS	IS2	.	>	N	^	n	¯			.	>	н	ч	Н	Ч
15	F	LAT	IS1	/	?	O	_	o	DEL			/	?	о	_	О	

CONTROLS

Figure 3b. USSR National Code



4 3 2 1		0 0 0	0 0 1	0 1 0	0 1 1	1 0 0	1 0 1	1 1 0	1 1 1	2 0 0	2 0 1	2 1 0	2 1 1	3 0 0	3 0 1	3 1 0	3 1 1
		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0 0 0 0	0	NUL	DLE	SP	0	@	P	`[1]	p				‘			’	±
0 0 0 1	1	SOH	DC1	!	1	A	Q	a	q			Ł	ł			‘	‹
0 0 1 0	2	STX	DC2	"	2	B	R	b	r			Ø	ø			´	›
0 0 1 1	3	ETX	DC3	#	3	C	S	c	s			Đ	đ			^	..
0 1 0 0	4	EOT	DC4	$	4	D	T	d	t			Þ	þ			~	°
0 1 0 1	5	ENQ	NAK	%	5	E	U	e	u			Æ	æ			¯	=
0 1 1 0	6	ACK	SYN	&	6	F	V	f	v			Œ	œ			˘	—
0 1 1 1	7	BEL	ETB	'	7	G	W	g	w			’	”			·	’
1 0 0 0	8	BS	CAN	(	8	H	X	h	x			·	ı			¨	‚
1 0 0 1	9	HT	EM	)	9	I	Y	i	y			ƀ	£			ˇ	“
1 0 1 0	10	LF	SUB	*	:	J	Z	j	z			®	ð			°	”
1 0 1 1	11	VT	ESC	+	;	K	[	k	{[2]			±				¸	„
1 1 0 0	12	FF	FS	,	<	L	\	l	¦			Ơ	ơ			‚	
1 1 0 1	13	CR	GS[3]	−	=	M	]	m	}[2]			Ư	ư			‚	
1 1 1 0	14	SO	RS[3]	.	>	N	^[1]	n	~[1]			’				”	’
1 1 1 1	15	SI	US[3]	/	?	O	_[1]	o	DEL							ˇ	

[1] Redefined elsewhere in the set.
[2] To be used as shift codes for 6-bit set (nonlocking).
[3] To be used as terminators or delimiters.

Figure 3c. USA National Code (ASCII) with library extensions

3.1 Addition of new graphic symbols. In recent years there has been considerable standards activity devoted to extending this code while retaining the philosophy that underlies its construction and control features. New control features and new sets of graphics have been introduced. The right hand side of Figure 3a is an example of new graphics [15]. This set includes the Katakana characters, i.e. a Japanese syllabic script. The set illustrates the principle: a new set of up to 94 graphics is introduced. Members of this set are invoked in the Japanese Industrial Standard 7-bit code by using an existing control, Shift-Out (SO). Shift-In (SI) restores the standard set. At the present time techniques are being developed to permit inclusion of larger sets of characters, such as that needed to represent (more specifically, to encipher) Kanji (literally, "Chinese") by using two standard graphics to represent a single Kanji character.

Figure 3b shows another language extension, that for the Cyrillic alphabet, taken from the national standard for the USSR [16]. This illustrates a useful principle. Wherever possible the corresponding Cyrillic and Latin letters occupy corresponding positions in the table. Thus a rough idea of a Cyrillic or Latin text can be obtained from the output of a machine that can print only one side of the table.

Figure 3c shows another example of extension. This introduces a set of graphics desired by librarians. The U.S. Library of Congress uses this code to distribute bibliographic information on magnetic tape to an international clientele [17]. Only 56 symbols are added. Diacritical marks are emphasized.

3.2 Addition of new controls. Means of extending the repertory of controls are also under development. The means anticipate the use of two techniques. Figure 3a, b and c illustrate one technique, that of adding to the "width" of a code to secure the 256 characters of an 8-bit code. In an 8-bit code columns 10 through 15 (excluding the positions 10/0 and 15/15) are reserved for graphic characters.* Columns 8 and 9 are controls. Thus an 8-bit code provides for the addition of 34 controls to those of the basic standard code.

---

*Here and in later sections of this paper, the positions in code tables are designated by coordinates: column/row

---

The other technique is the use of a special code extension character, the meaning of which depends on how it is used. In the 7-bit code this is Escape (ESC). Escape is used as the first character of a sequence of two or more characters which, as a unit, represent a single control function.

The doctrines being developed for code extension provide for various additional levels of complexity. Our concern is only with the first level of extension, where the doctrines are already being implemented in hardware and software. For example, the important functions of Half Line Feed Forward, Half Line Feed Reverse, Reverse Line Feed, Clear Horizontal Tabulation Stops and Set Horizontal Tabulation Stops are available on stock teletypewriters [18] where they are represented by two-character Escape sequences. None of these controls is in the basic ISO set.

3.3 Code extension - general remarks. Current standardization work on code extension envisions a family of 8-bit codes and extended 7-bit codes, in which each member of the family retains the facilities of the basic 7-bit code as a subset [13]. (This is similar to the treatment of the keypunch and the common computer printer. They are contained in and defined as a subset of the basic 7-bit code [19]).

This concept of adding levels of complexity while retaining standard subsets is crucial to the orderly development of general purpose facilities over a long period of time. It is used today to assure a high level of correspondence among national codes. For example, ASCII (columns 2 through 7 in Figure 3c) differs from the Japanese standard code (same columns in Figure 3a) at only three positions: 5/12, 7/12, and 7/14. The variations in Figure 3b also are minor. Thus the basic ability to communicate among systems supporting the three codes shown (figures 3a-c) is assured. For example, virtually all of the text of this paper could be printed by any of them.

The examples of code extension discussed above emphasize enlargement of the graphic character repertory. Clamons [20] has published a summary of current work on character codes which includes a summary of proposals for extensions of the control repertory with some emphasis upon extensions intended to be useful in cathode ray display devices. Clamons' paper includes a multicolored chart which provides a very convenient summary of the inter-relationships among ASCII structured character codes.

4. A General Purpose Scientific Document Code (GPSDC)

Code extension, discussed in Section 3, has been used to adapt the information interchange standard to the needs of science. In effect, this converts the 7-bit code to an 8-bit code and provides for up to 34 new controls and 94 new graphic characters.

4.1 The extended code. The set of characters proposed here for use in documentation of scientific work is shown in Figure 4. This proposal is a substantial revision of that originally made [7] and is slightly different from the set of graphics initially selected for installation on a printer at NBS [8]. The revisions and rearrangements have been made on the basis of experience with the earlier set, certain standards developments and comments received. They have been made in the hope that they will increase the utility of the code.

These 189 characters form a set of "primitive graphics" that are to be realized in hardware. The needs of science are not met completely by these. They must be supplemented by the techniques described in the following three subsections.

The best approximation of this set realized to date is shown in Figure 5. This is a reproduction of the actual computer printed code table used in documentation describing the facilities of the NBS Computer Services Center used to produce

---

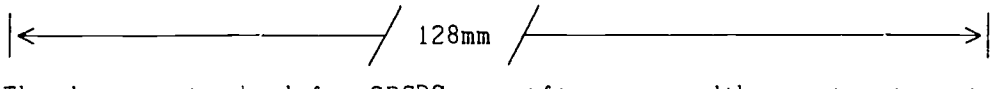
Figure 4

General Purpose Scientific Document Code. This is an extension of the ISO Code to meet the needs of scientific text. Columns 0-7 have the basic 7-bit code and columns 8-15 the extension. Compare with Figures 3a, 3b and 3c. The new controls, in columns 8-9, are:

NL	New Line
ECP	Execute Control with Parameter
RLF	Reverse Line Feed
HLR	Half Line feed Reverse
HLF	Half Line feed Forward
ESO	Extended Shift Out
ESI	Extended Shift In
ACP	Accept Control Parameter
AGQ	Accept Graphic Qualifier
TCG	Terminate Composite Graphic
DT1	Device Test 1
DT2	Device Test 2
DT3	Device Test 3
r*	reserved for future assignment

GPSDC Code Table
Defining Representation - 1972 November 10

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	NUL	DLE	SP	0	@	P	`	p	r*	r*	□	•	∂	Π	°	π
1	SØH	DC1	!	1	A	Q	a	q	r*	r*	'	▪	◊	θ	α	θ
2	STX	DC2	"	2	B	R	b	r	r*	r*	⌒	▪	/	√	ß	ρ
3	ETX	DC3	#	3	C	S	c	s	r*	r*	£	▪	/	Σ	∫	σ
4	EØT	DC4	$	4	D	T	d	t	r*	r*	¤	▪	Δ	↑	δ	τ
5	ENQ	NAK	%	5	E	U	e	u	r*	r*	∝	⟍	∃	Υ	ϵ	υ
6	ACK	SYN	&	6	F	V	f	v	r*	r*	∞	⟍	ϕ	↓	ϕ	∇
7	BEL	ETB	'	7	G	W	g	w	r*	r*	|	⟋	Γ	Ω	γ	ω
8	BS	CAN	(	8	H	X	h	x	r*	r*	⊂	⟋	\	∩	η	χ
9	HT	EM	)	9	I	Y	i	y	NL	r*	⊃	|	\	Ψ	ι	ψ
10	LF	SUB	*	:	J	Z	j	z	ECP	ACP	×	≠	Ξ	U	ξ	ζ
11	VT	ESC	+	;	K	[	k	{	RLF	AGQ	~	-	-	⫽	κ	§
12	FF	FS	,	<	L	\	l	|	HLR	TCG	_	←	Λ	⟍	λ	|
13	CR	GS	-	=	M	]	m	}	HLF	DT1	¬	≈	••	⫽	μ	¶
14	SØ	RS	.	>	N	^	n	~	ESØ	DT2	—	→	||	⌣	ν	-
15	SI	US	/	?	Ø	_	o	DEL	ESI	DT3	⟋	▮	¤	⌣	∘	EØ



The design standard for GPSDC specifies monowidth spacing in metric units for the basic set of 189 typewriter-like printing characters. Character depth (vertical line spacing) is 4mm and character set width (horizontal spacing) is 2mm. The defining representation implies digitalization on a dot matrix with dots spaced 0.05mm on centers and printing using dots with a fixed diameter lying between 0.1 and 0.15mm. The characters slant, 2/15, and long dash, 10/14, are, respectively, a full diagonal and a full horizontal stroke in the print window used to define the basic set.

GPSDC Code Table†

		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
		0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	0	NUL	DLE	SP	0	@	P	`	p	r*	r*	□	•	∂	Π	•	π
1	1	SØH	DC1	!	1	A	Q	a	q	r*	r*	′	·	◊	the	α	∂
2	2	STX	DC2	"	2	B	R	b	r	r*	r*	bc	•	ob5	✓	β	ρ
3	3	ETX	DC3	#	3	C	S	c	s	r*	r*	pst	•	ob6	Σ	∫	σ
4	4	EØT	DC4	$	4	D	T	d	t	r*	r*	csm	•	Δ	↑	δ	τ
5	5	ENQ	NAK	%	5	E	U	e	u	r*	r*	∝	ob1	∃	T	ϵ	υ
6	6	ACK	SYN	&	6	F	V	f	v	r*	r*	∞	ob2	Phi	↓	phi	∇
7	7	BEL	ETB	aps	7	G	W	g	w	r*	r*	'	ob3	Γ	Ω	γ	ω
8	8	BS	CAN	(	8	H	X	h	x	r*	r*	(	ob4	ob7	∩	η	χ
9	9	HT	EM	)	9	I	Y	i	y	r*	r*	)	'	ob8	Ψ	ι	ψ
10	A	LF	SUB	*	:	J	Z	j	z	ECP	ACP	×	⇌	Ξ	∪	ξ	ζ
11	B	VT	ESC	+	;	K	[	k	{	RLF	AGQ	∽	-	-	≈	κ	§
12	C	FF	FS	,	<	L	\	l	|	HLR	TCG	-	←	Λ	\	λ	|
13	D	CR	GS	-	=	M	]	m	}	HLF	DT1	¬	≈	∞	≠	μ	¶
14	E	SØ	PS	.	>	N	^	n	~	ESØ	DT2	—	→	∥	˘	ν	¯
15	F	SI	US	/	?	Ø	_	o	DEL	ESI	DT3	∕	▮	∝	bf	ο	EØ

†Rendition set up for IBM 1416 Train Cartridge, Order # G 83154-55, for use on an IBM 1403-N1 Printer with features: (1) Universal Character Set, (2) Wide Hammers and (3) 16 Lines per Inch Spacing. See: NBS Tech. News Bull. 54 No. 2, 35 (Feb. 1970)

□, MSP, "meta space", use restricted, ESC 3/0 in 7-bits

EØ, "Eight Ones", use restricted, ESC 3/15 in 7-bits

r*, controls reserved for future assignment

Graphics not on this printer

aps, apostrophe, 10/7 prints instead
bc, bracket cap
bf, bracket foot
csm, currency sign
ob1 - ob8, "octobliques", See: Gottardi, J. Chem. Doc. 10 75 (1970)

Graphics not on this printer as simple symbols but that can be constructed

Phi, Greek Φ
phi, Greek ϕ
pst, pound sterling sign, £
the, Greek θ

Figure 5. Approximation of the GPSDC table using an existing high speed line printer.

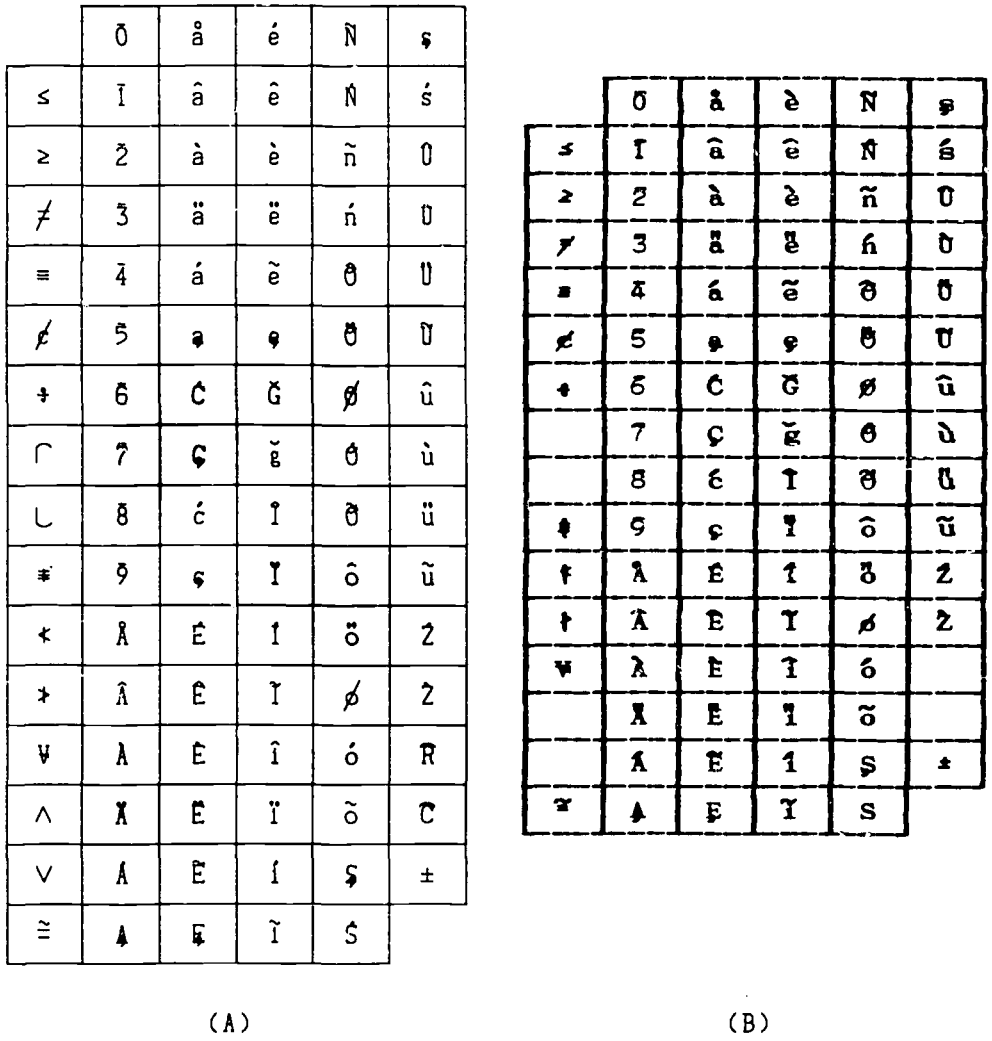
the camera-ready copy for this Technical Note. The character at position 10/0, Meta Space, is considered a control not freely available to the user as a notational device. The character is discussed further in Section 4.9.

4.2 Internal Extension - Composite Symbols. Figure 6 displays examples of "composite" characters rendered by two different printing mechanisms. Composites are important because they provide an "internal extension" to obtain additional characters.

Composites are overstrike combinations of two symbols, such as combining a "less than" (<) and "low dash" (-) to make "less than or equal" (≤). This technique is commonly used for accented letters. Here it is exploited more fully. Any pair of symbols can be combined. The combination should have an easily recognized meaning. The basic ISO code document allows for the use of this technique of combining two symbols in one location to create a symbol with a different meaning. "Backspace" is commonly used in this technique. Figure 6 shows only a partial listing of the composite symbols presently being used in GPSDC. By implication it indicates our practice of cataloging composites in sets of 94 to allow for possible future developments which would treat our composite symbols as primitive symbols in alternative sets of graphic characters to be invoked using higher level code extension techniques [13].

4.3 Class Modifications. Another "internal extension" is class modification. Seven class modifications are specified. These are produced by underscoring and/or overscoring with dashes, waves, arrows and dots. These modifications apply to all the symbols in the set, including the composites. The meaning of a particular modification is not defined. However, the most common use to date has been to indicate several type faces. Class modification is the typescript notation which provides the implementation of the stylistic requirements, i.e. italic, bold face, etc., of the IUPAC rules cited in section 2.

4.4 Control Functions. The addition of two control functions completes this code for science. These controls are "Half Line Feed Forward" and "Half Line Feed Reverse". These are vertical motions on a page. They permit placement of superscripts and subscripts. As previously described, these have already been introduced by some manufacturers in the first level of extension of the control set for the standard code by use of escape sequences.



(A) (B)

FIGURE 6

Composite symbols. Examples of new characters formed by overstriking pairs of characters from Figures 4 and 5.

a. Composites drawn by a computer driven incremental plotter (see Section 4.8) using the preferred 2:1 print window, Figure 4 characters.

b. Composites as produced on the NBS line printer showing the distortion caused by adoption of a 1.25:1 print window, Figure 5 characters.

4.5 Selection of graphic symbols. The general basis for symbol selection has been described in section 2. The required techniques of using composite symbols and class modification are explained earlier in section 4. A detailed defense of each symbol is inappropriate here. The objective is to provide a set in which there exists one form for each required symbol, not the entire range of alternatives that could be used. It is expected that the scientist who uses this system will have to make some compromises. He does this now: he must work within the set of symbols available to his typist. We are confident that the code provides reasonable solutions to all but the most abstruse problems of preparing a typescript.

4.6 Options in the ISO Code and in ASCII. Practical experience suggests that a graphic character set can be selected which will be found to be serviceable in non-scientific applications, or at least in technical applications not originally contemplated. Our original emphasis on recording scientific text may be unwarranted. This possibility is important to our computer center and probably to others.

For this reason, certain non-scientific symbols have been included. The approach has been to incorporate symbols that may have alternative forms in the ISO and ASCII codes. Thus, the extension includes the Pound Sterling and the General Currency Symbol which are, in the ISO code, proposed alternatives for the Number Sign (#) and the Dollar Sign ($) in ASCII. A solid Vertical Line has been substituted for the broken Vertical Line at 7/12 (Fig. 3c) in anticipation of a revision of ASCII to correspond to international practice.

4.7 Diagrams. A more important consideration is the provision of an adequate set of rule segments for diagrams and chemical structures. The set proposed by Gottardi [21] for chemical structures has been included in toto. This set is, in our opinion, equally useful for many classes of diagrams. It should be considered very seriously by equipment designers. The set of rule segments and plotting dots is collected in columns 4′ and 5′ of the Shift Out set in Figure 7. Taken alone they form a reasonable extension of the basic ASCII set and could be realized on a Model 37 "Teletype" (which can print 126 symbols).

Shift In Set

	0	1	2	3	4	5	6	7
0	NUI (DEL)	DLE (DEL)	SP	0	@	P	`	p
1	SOH (DEL)	DC1 (DEL)	!	1	A	Q	a	q
2	STX (DEL)	DC2 (DEL)	"	2	B	R	b	r
3	ETX (DEL)	DC3 (DEL)	#	3	C	S	c	s
4	EOT (DEL)	DC4 (DEL)	$	4	D	T	d	t
5	ENQ (DEL)	NAK (DEL)	%	5	E	U	e	u
6	ACK (DEL)	SYN (DEL)	&	6	F	V	f	v
7	BEL	ETB (DEL)	'	7	G	W	g	w
8	BS	CAN (DEL)	(	8	H	X	h	x
9	HT	EM (DEL)	)	9	I	Y	i	y
10	LF	SUB (DEL)	*	:	J	Z	j	z
11	VT	ESC	+	;	K	[	k	{
12	FF	FS (DEL)	,	<	L	\	l	|
13	CR	GS (DEL)	-	=	M	]	m	}
14	SO	RS (DEL)	.	>	N	^	n	~
15	SI	US (DEL)	/	?	O	_	o	DEL

Shift Out Set

	2'	3'	4'	5'	6'	7'
0		0	-	'	(DEL)	(DEL)
1	!	1	~	||	(DEL)	(DEL)
2	"	2	—	\	(DEL)	(DEL)
3	#	3	•	•	(DEL)	(DEL)
4	$	4	•	\	(DEL)	(DEL)
5	%	5	|	\	(DEL)	(DEL)
6	&	6	/	•	(DEL)	(DEL)
7	'	7	/	//	(DEL)	(DEL)
8	(	8	\	•	(DEL)	(DEL)
9	)	9	/	\	(DEL)	(DEL)
10	*	:	\	-	(DEL)	(DEL)
11	+	;	n	'	(DEL)	(DEL)
12	,	<	'	-	(DEL)	(DEL)
13	-	=	u	'	(DEL)	(DEL)
14	.	>	⌢	≠	(DEL)	(DEL)
15	/	?	\	⌣	(DEL)	

'DEL) indicates 'no page printer response', i.e. equivalent to Delete.

4.8 Proposed Standard Print Window. The design of a set of graphics suitable for the formation of composite symbols and for the construction of diagrams requires the consideration of compatibility in connection with almost every member of the set. In the course of this work it became clear that we needed to adopt some concept of a standard "print window" realizable in hardware. A print window is the rectangular space within which a graphic character is placed. For mono-width character sets of the same size or font all symbols have the same print window, although many of them occupy only a portion of it. (Indeed, the space around a symbol is an important element in its design.) For typewriting on single spacing, the print windows fill a page. For two adjacent windows on a line the left edge of the second is the right edge of the first. For two adjacent windows in a column the top edge of the lower is the bottom edge of the upper. When subscripts and superscripts are produced by half-line spacing, print windows may overlap vertically.

At present, the existing hardware (line printers, displays, typewriters) shows considerable variation in character and interline spacing. No one print window is applicable to all. Different compromises must be made in the design of a set of graphics. For example, existing computer line printers employ the spacing used on the relatively uncommon "pica" typewriters where the horizontal spacing is 10 characters per inch and the vertical spacing is 6 lines per inch. In contrast, Gottardi achieved rational slopes for rule segments by having his printer modified to space 10 half lines per inch. In his implementation a square print window is used.

Our recommendation is neither of these. It is a print window with an aspect ratio of 2:1. It is further specified that graphic characters may extend over the entire height and width of the print window, although

---

FIGURE 7

An extended code set. One hundred twenty-six distinct characters. Columns 2′ and 3′ repeat columns 2 and 3. The extension in columns 4′ and 5′ emphasize plotting dots and rule segments. Figure produced directly on a Model 37 "Teletype".

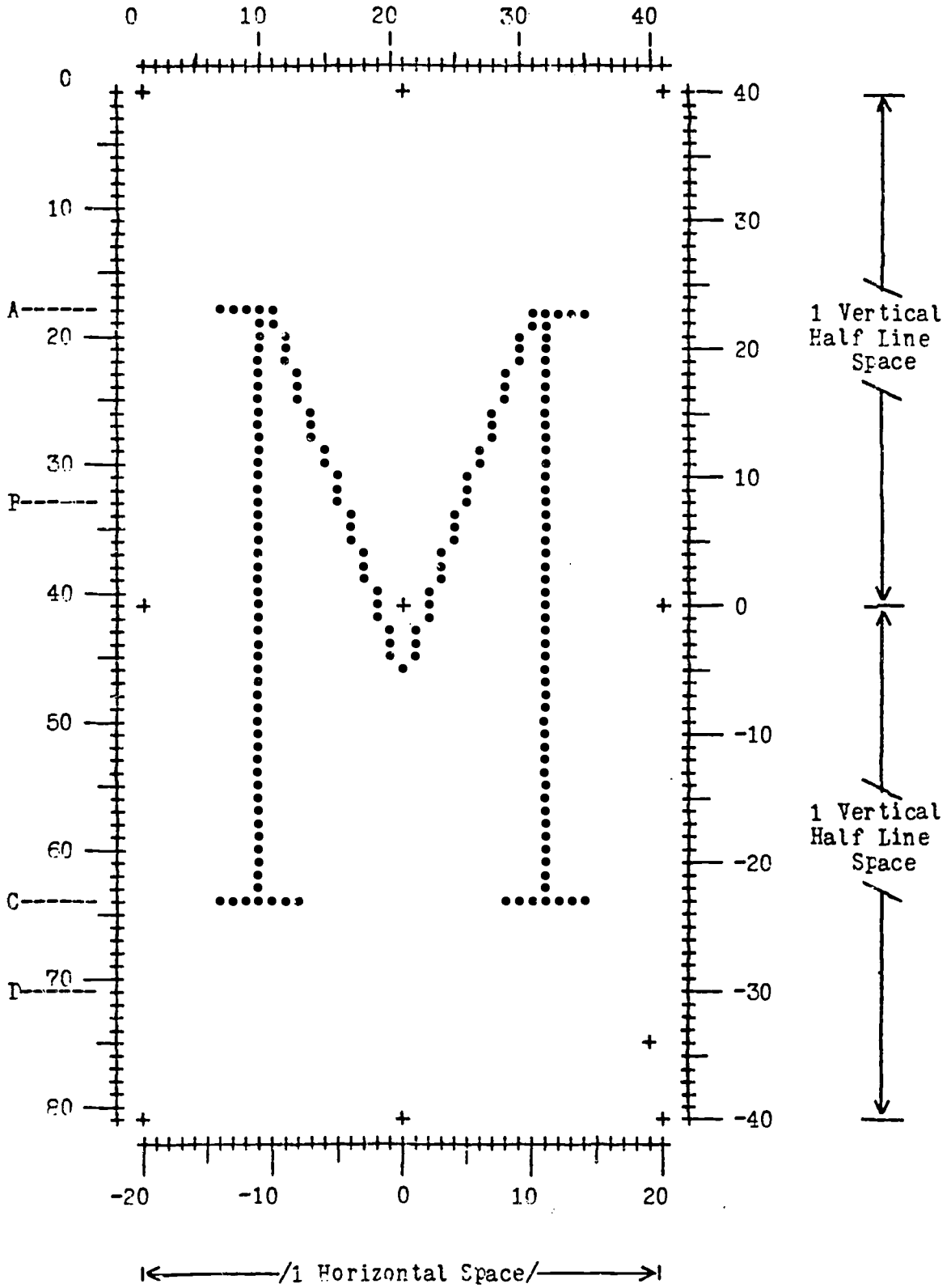


Figure 8. Binary specification of a symbol in the standard print window.

most of them do not. This 2:1 print window meets the requirements for readable mono-width characters: It has the same relative spacing as the common "elite" typewriter (12 characters per inch horizontally, and 6 lines or 12 half lines per inch vertically). It also permits the use of the set of rule segments specified by Gottardi.

On typewriters one can secure printing over a full print window so specified. However, this is not possible on any existing computer line printer we have examined. For example, in the case of the printer used to produce the copy for Figure 5 and the bulk of the copy for this Note the maximum permitted vertical extent of a graphic character is 0.137 inch. Thus, in adapting the GPSDC graphic character set to this printer we introduced distortion by using a print window appropriate to 8 lines per inch vertical spacing (1/16 inch half line space) and 10 characters per inch horizontal spacing. The aspect ratio is 1.25:1. The resulting distortion is of no particular consequence in ordinary text. However, a comparison of the two renderings of composite symbols in Figure 6 shows that cramping in the vertical direction has led to printed composite symbols which are not quite satisfactory. In addition, in carrying out the distortion of symbols we altered the slopes of the rule segments and the positions of the plotting dots.

Figure 2 suggests that this distortion will be of little consequence in rendering some classes of diagrams. However, it will be important in scale drawings such as Figure 8. We feel that this compromise with existing hardware is a temporary expedient. The dimensional specifications of line printers can be changed in future models if manufacturers so choose. In any case alternatives to the common line printer mechanism are available and are coming into more common use.

A part of our current program of work is the implementation of a convenient scheme for simulating a document writer using a computer driven incremental plotter. The driver programs make use of a catalog of digitalized graphic characters. This effort is intended to produce a design tool which can produce individual large scale drawings of graphics of the kind suitable for documenting specifications while, at the same time, being able to produce smaller scale text so as to test the compatibility of the graphics. In planning this portion of the work we have anticipated certain advantages in having a catalog of symbol specifications which can be rendered on a document writer providing

only the printing facilities of GPSDC. Thus, we have chosen to develop binary specifications of symbols as illustrated in Figure 8. We describe these specifications as "binary" because the symbol specification is rendered at large scale by a dot - no dot scheme using the large centered dot at position 11/0 in the code table.

This part of the work is in large measure an experiment to help us better understand the implications of "registration" of coded character sets. The notion of "registration" as an element of the standardization process has been invoked at the USA Federal [22] and the ISO [23] levels. This whole paper, not only the particular facility being discussed at this point, speaks to the mechanisms of analysis, evaluation and documentation of the kind with which, it seems to us, any "registration authority" would be concerned.

It is not possible here to give an account of all the background considerations which led to the symbol cataloging scheme chosen. Basically, we chose a compromise working scheme which gives definitions somewhat more coarse than that required to carry the full artistic burden of graphic arts typography [24,25]. On the other hand the specification is judged to be sufficiently fine grained to carry those elements of distinctiveness required for symbol recognition on the part any human reading for content. In addition, when prepared on a typewriter, Figure 8 has a scale of approximately 40:1. We have prepared input specification forms at this same scale; it is a convenient one with which to work.

At this time it does not appear to be possible to define a "best" choice for the absolute dimensions of Figure 8. For the time being we are suggesting that the horizontal single space shown in Figure 8 be taken 2 millimeters. This suggestion takes into account the effort to progress toward a metric scheme for measuring typefaces as exemplified by a recently adopted British standard [26]. At present we are cataloging what could be described as typewriter style graphic characters designed for monowidth spacing where the single line spacing ("character depth" in the typographic standard cited) is twice the character width. In Figure 8 the marks labeled A, B, C and D indicate, respectively, the tops of large letters, the tops of small letters without ascenders, the base line for letters and the bottom of descenders on those letters where they occur.

4.9 The Meta Space Character. When a user interacts with a system through the interchange of printed text, there is almost invariably a need to mix data text with control text. There are many uses for visible flags and delimiters. There is sometimes contention between systems designers and users over the matter of reserved symbols. The character Meta Space was introduced into the GPSDC character set to aid in alleviating the problem of reserved symbols. System implementors are encouraged to consider the following guidelines in devising uses for Meta Space.

(1) When Meta Space occurs at a print position by itself, i.e. not as part of a composite, it may be treated as ordinary space or, perhaps, "required space" to indicate that a textual element including space is not to be broken in adjusting line lengths. In manuscripts for processing by a typesetting program it may be used to represent "em space".

(2) When the need arises to define a visible control or delimiter, Meta Space should be one of the components of a composite symbol assigned the required meaning. For example, in one processing system used at NBS, Meta Space plus Exclamation Point is used as a Command Prefix. In this same system Meta Space overstruck with a Question Mark is used as a location-specific diagnostic error signal in system generated printing.

(3) Output options should be used to control the printing or suppression of Meta Space, composite symbols involving Meta Space or commands delimited through the use of Meta Space.

(4) The system designer must specify each proper use of the Meta Space character. (188 printing characters are available for recording data, only this one is reserved for control purposes.)

4.10 Keystrokes versus Characters. In designing GPSDC we have taken into account the developments which are leading to economical solid-state logical components associated with keyboards. We think, for example, keyboard operators will not continue to have to use two keystrokes to represent a

"New Line" through the successive actions of Line Feed (LF) and Carriage Return (CR). We expect modern keyboards which emit standard codes to be augmented by character sequence generators which will emit the proper sequences of characters to represent frequently used functions. As a consequence we have not, in general, equated "number of characters" with "number of keystrokes" in deciding which control functions or symbols should be assigned as single characters in the GPSDC code table.

5. Preparation of copy and input devices

The input process and the desirable characteristics of input machines are described in this section. The discussion is limited to the preparation of copy on typewriters. We have not made a sufficient examination of cathode ray display devices.

5.1 Preparation of copy - the operator's view. The basic mode of capture of text should be as similar to ordinary typewriting as possible. The operator should be a scientific typist, not a compositor. This reduces the need for special training.

Free-form copy must be acceptable. It should be possible to produce a page of copy that is exactly the same as the final copy of the typescript of a scientific paper. When an ideal machine is used this means that all symbols will appear on the page clearly readable, i.e., with no ciphers, in their proper positions, and without the introduction of visible control information.

An input system may allow for highly stylized typing of fragments of copy followed immediately by reformatting. This is very useful in an interactive system. It should be an addition to, not a substitute for, the basic free-form input mode.

5.2 Preparation of copy - machine processing requirements. It is assumed that the record produced by the input device will be processed by a computer program. This is desirable because the process can eliminate many restrictions on the typing. Records can be cleaned up to correct errors recognized by the typist. Also the string of codes representing a line of text can be put into a standardized sequence that will simplify later processing, particularly information retrieval. Computer processing will be necessary if the

input machine produces a code sequence other than ISØ. This translation should be accepted as a normal procedure.

It is mandatory that each keyboard action be encoded in the record produced. These keyboard actions should be sufficient to produce the full record.

5.3 Specific Features of typewriter-like input machines. The important features for an input device are:

(1) A basic set of 80-94 graphic characters and the normal set of controls that affect the position of text on a page, e.g. line feed, carriage return, space and tabulate.

(2) Øverstrike capability, e.g. backspace.

(3) Half line feed forward (down) and Half line feed reverse (up).

(4) Access to an alternative set of graphic characters and control codes for them.

These features are listed in order of their importance. In almost every case selection of an input device will require some compromise. There is no device that matches GPSDC perfectly, but good approximations can be found.

The basic set of graphic characters. That set normally supplied by manufacturers is acceptable but usually inadequate. If a choice is available, maximize the number of those characters used frequently or those in the GPSDC set.

Øverstrike capability. This is useful for three purposes: correction of errors (by substitution), class modification and construction of composites. Since these composites may be ciphers for non-available characters, overstriking is a very powerful method for extending a basic set of characters. Both corrections and composites are to be interpreted by the processing program. Whenever possible, the input keyboard operator should not be required to observe a prescribed sequence of operations in overstriking.

Half line feed controls. These permit clean encoding of superscripts and subscripts and make the entire set of characters available in these positions. They are also needed for diagrams. If a machine does not have these controls,

super- and subscripts must either be ciphered as composites, or be included in the basic set. Both of these choices are poor.

Alternative set of graphics. Access to a second set of graphics permits clean text preparation for more complex material. At times substitution of a second set is easy, e.g. when typespheres are used. But even when the physical act is difficult, the controls to invoke the alternate set of graphics are useful. They should be of the type "Select set X", i.e. shift and lock, or "Select the next character from set X", i.e. non-locking shift. The use of a single control to alternate between two sets of graphics should be avoided. The operator has no reliable method for restoring a basic condition in case of error.

The choice of the alternative set should be made to maximize the number of GPSDC characters.

5.4 Experience at NBS. The data centers that use the GPSDC system have a variety of recording typewriters. They differ considerably in their features and ease of operation. The machines in use are identified below. The specific features (section 5.3) that each has is shown in braces. Friden "Flexowriter" (circa 1961), {1, 2}, SCM (CDC) "Typetronic" {1, 2}, "Taxywriter" {1, 2, 4}, Dura (Itel) Model 1041 {1, 2, 3, 4}. IBM "MTST" {1, 2, 3, 4?}, IBM "MCST" {1, 2, 3, 4?}, "Model 37 "Teletype" {1, 2, 3, 4}. Other models of these machines may have more features. We do not expect to have difficulty in handling input from any typewriter that records on a standard medium.

Selection of sets of graphic symbols has proved easiest for the typewriters that use typespheres, e.g., the "Selectric" typewriters, and for the "Teletypes". This is simply because the former provides 150-160 useful symbols with two stock spheres and the latter can provide 126 characters in its typebox.

Explicit control features for selecting alternative graphic sets are not available on most of the machines mentioned above. The practice has been to devote some unused peripheral control function to this purpose. "Red ribbon shift" is ideal for this purpose and has been used to advantage on "Teletypes" to indicate Greek, etc.

Overstriking, particularly for corrections, is so widely used that a machine without this capability would be unacceptable in the data centers.

The use of keypunches to produce scientific text must be mentioned. A complex coding scheme that makes a keypunch simulate a typewriter {1, 2, 3, 4} was developed as an emergency measure. To our surprise, this has been used extensively, particularly in the editing of records.

6. Use of the GPSDC System at NBS

6.1 Implementation. A text-handling system based on the General Purpose Scientific Document Code has been in operation since 1967. It is an experimental system. There has been a constant need to refine the techniques used and to improve the definitions of the system. The experimental nature of the system has not daunted its users. They have employed it for day to day production, and often have used it for tasks not originally contemplated. They have written substantial programs for special applications.

The programs for the system are written in FORTRAN. They are used in batch mode on the NBS Univac 1108 under EXEC II and EXEC 8 and previously ran on a CDC 3100. The program deck totals about 25,000 cards. Batch mode is employed, not out of preference but because only it is available. Many of the programs would be applicable to an interactive mode.

The basic system provides for input from recording typewriters, keypunches and from magnetic tapes prepared either in ASCII 1968 or by on-line text processing systems, such as the IBM ATS. There are programs for editing, reformatting, search and retrieval and for output to a line printer, a photocomposition machine and to ASCII 1968 on magnetic tape.

6.2 Users and extent of use. Three data centers regularly use GPSDC to prepare their files of information about their specialties; thermochemistry, chemical kinetics and atomic spectroscopy. Other groups have used the system to prepare books. These users take advantage of two features: access to a line printer for proof copy and access to a high speed photo-composition machine. The subject disciplines are statistics, diatomic spectra, molecular structure, analytical chemistry, radiation chemistry and crystal structure.

The total usage of the system is small when measured against other automation. During the 12 months ending June 1970, about 100 jobs per month were logged in by the NBS Computer Services Center. The current (August 1973) rate of utilization is about 150 jobs per month. This usage reflects the size of the centers (one to ten persons) and the work load they can handle.

6.3 Typical applications. The GPSDC system was designed for recording, in machine-readable form, typescript records that must be saved and referred to, but will not be published. Data evaluation groups and information centers produce such records in large quantities as they scan the literature in their fields, select some articles for retention and then abstract and index them. Recording this information is the typical use of the GPSDC system. An example is shown in Figure 9. Because of the technical orientation of the centers, scientific notation is mandatory.

There is a much larger class of records produced by any institution: file copies of administrative memoranda and actions. A simpler script and recording convention usually is sufficient. But administrative and technical records merge in documents such as annual reports on research accomplishments, written for managers but inevitably containing technical notation. Circumlocutions and awkward `spelling out' of symbols are characteristic marks of these documents (and of abstracts) written to fit within the limits of the simpler scripts.

6.4 Special applications. Several of these warrant detailed description. They are evidence supporting our contention that the printed scientific document can be prepared in GPSDC. The Bulletin of Thermodynamics and Thermochemistry [27] is prepared using the GPSDC system. Three groups (two outside NBS) abstract and index current articles using a highly stylized form. These records are sent to NBS on punched paper tape, converted to GPSDC, returned for proofreading and then edited. The records are processed for publication by programs that construct a bibliographic section and an index arranged by chemical formulas. (The programs interpret the fomulas written in normal scientific notation and assign the indexing sequence.) Since 1971 four sections of the Bulletin (organic substances, organic mixtures, inorganic substances and bibliography) have been photocopied from output from the NBS line printer. The 1970 inorganic substances section also was prepared in this way. In 1969, GPSDC records for this section were printed via a tape driven typewriter. A magnetic tape version of the 1971 Bulletin, coded in ASCII 1968, has been produced from GPSDC records and issued as NBS Magnetic Tape No. 4 [28].

Several books and journal articles have been produced. "Tables of Molecular Vibration Frequencies", NSRDS-NBS 39 [29], was keypunched at the University of Tokyo and the cards sent to NBS to be processed into GPSDC. Line printer copy was

BRIEF: FW/4569 Page 1 CHPLB-1968-2-143 CKIC/15403

AUTH: Brennen, W., and Shane, E. C.

TITLE: Pressure-Dependence of the Yellow Nitrogen Afterglow Intensity

REF: Chem. Phys. Letters (Amsterdam) 1968 2 143

REACT: $N + N + M \rightarrow N_2(A^3\Sigma_u^+) + M$

$N_2(A^3\Sigma_u^+) + M \rightarrow N_2(B^3\Pi_g) + M$

$N_2(B^3\Pi_g) + M \rightarrow N_2 + M$

$M = N_2$

$N_2(B^3\Pi_g) \rightarrow N_2 + h\nu$

$N_2(A^3\Sigma_u^+) + M \rightarrow N_2 + M$

INDEX: Experimental: gas: Bond cleaved NN: Bond formation NN: pressure: energy-transfer: fluorescence: excitation: quenching: rate: raditive: electronic: second-order: chemiluminescence: nitrogen-molecule (product):

(a)

FIGURE 9

Data Center Records. Indexing record prepared for the Chemical Kinetics Information Center, NBS.

a. Input copy prepared on a Model 37 "Teletype". Ciphers (overstrike combinations) are used to encode Greek letters.

b. Line printer copy of the same record. The ciphers have been interpreted.

Book 1 Page 3 *...30...35...40...45...50...55...60- y Ln eedd

BRIEF: FW/4569 Page 1 CHPLB-1968-2-143 CKIC/15403 - 2 1

AUTH: Brennen, W., and Shane, E. C. - 5 2

TITLE: Pressure-Dependence of the Yellow Nitrogen - 8 3
Afterglow Intensity - 11 4

REF: Chem. Phys. Letters (Amsterdam) 1968 2 143 - 14 5

REACT: $\tilde{N} + N + M \rightarrow N_2(A^3\Sigma_u^+) + M$ - 18 6

$N_2(A^3\Sigma_u^+) + M \rightarrow N_2(B^3\pi_g) + M$ - 22 7

$N_2(B^3\pi_g) + M \rightarrow N_2 + M$ - 26 8

$M = N_2$ - 29 9

$N_2(B^3\pi_g) \rightarrow N_2 + h\nu$ - 33 10

$N_2(A^3\Sigma_u^+) + M \rightarrow N_2 + M$ - 37 11

INDEX: Experimental: gas: Bond cleaved NN: - 40 12

Bond formation NN: pressure: energy-transfer: - 43 13

fluorescence: excitation: quenching: rate: - 46 14

radiative: electronic: second-order: - 49 15

chemiluminescence: nitrogen-molecule (product): - 52 16

Figure 9b

proofread in Tokyo and then returned for correction of the GPSDC file and preparation for phototypesetting. An example of the final output is shown in Figure 10. A similar example, Figure 11, is the preparation of tables of data on rates of reaction of electrons in water solution [30]. The input was from formatted typed tables (on paper tape). The figure demonstrates successful conversion to printing in which there are stylistic complexities. These examples and others and the techniques used are presented in more detail in reference [31].

The multivolume handbook "Crystal Data" [32] has been prepared for printing by using this system in a special manner. The original copy was keyboarded for one photocomposition machine (Mergenthaler Linofilm). Later a decision was made to use another machine (Mergenthaler Linotron) because a better printing and publication schedule could be obtained. Linofilm records are converted to GPSDC, proof copies made on the line printer and then the corrected copy processed to drive the Linotron.

Conversion of other records from Linofilm coding has been used to advantage by the Chemical Kinetics Information Center. Several monographs on kinetics were keyboarded by the U.S. Government Printing Office as part of the normal publication process [33]. The machine-records were converted to GPSDC in order to add the contents of these handbooks to the magnetic tape file on kinetics. Proportional spacing is lost in the conversion, but virtually all symbols used by the compositors were converted properly. Probably, Monotype records could be converted in a similar manner.

Two other examples of the use of "foreign" machine records should be mentioned. Memoranda, technical typescripts and bibliographies prepared using on-line text processing systems are converted to GPSDC for long term storage. The American Institute of Physics SPIN tapes, a current awareness service, are searched at NBS for several data centers. These tapes are ciphered to indicate upper and lower case, subscripts and superscripts and use names for Greek and special characters. The retrieved material is converted in GPSDC, reformatted to match the desires of each center and printed in clear text.

6.5 One system for many users. All text handling tasks are very similar in the demands that they make on a system. This warrants a general approach in which modules of considerable flexibility are invoked at the various stages of input, editing, reformatting, retrieval and printing. Flexibility is important. By this is meant ability to handle a large class of closely related variants of the same task. The special

Molecule: Chloroacetylene HCCCl **No. 148**
Symmetry $C_{\infty v}$ Symmetry number $\sigma = 1$

Sym. class	No.	Approximate type of mode	Selected value of frequency	Infrared	Raman	Comments
				cm^{-1} (Gas)	cm^{-1}	
σ^+	ν_1	CH stretch.....	3340 B	3340 VS		
	ν_2	C ≡ C stretch.....	2110 B	2110 VS		
	ν_3	CCl stretch.....	756 B	756 VS		
π	ν_4	CCH deg. deform.....	604 B	604 S		
	ν_5	CCCl deg. deform.....	326 B	326 W		

References

[1] IR. W. J. Middleton and W. H. Sharkey, J. Am. Chem. Soc. **81**, 803 (1959).
[2] IR. W. S. Richardson and J. H. Goldstein, J. Chem. Phys. **18**, 1314 (1960).
[3] IR. G. R. Hund and M. K. Wilson, J. Chem. Phys. **34**, 1301 (1961).

Molecule: Bromoacetylene HCCBr **No. 149**
Symmetry $C_{\infty v}$ Symmetry number $\sigma = 1$

Sym. class	No.	Approximate type of mode	Selected value of frequency	Infrared	Raman	Comments
				cm^{-1} (Gas)	cm^{-1}	
σ^+	ν_1	CH stretch.....	3325 B	3325 VS		
	ν_2	C ≡ C stretch.....	2085 B	2085 VS		
	ν_3	CBr stretch.....	618 C	618 VS	.....	SF (ν_4).
π	ν_4	CCH deg. deform.....	618 C	618 VS	.....	SF (ν_3).
	ν_5	CCBr deg. deform.....	295 B	295 W		

References

[1] IR. W. J. Middleton and W. H. Sharkey, J. Am. Chem. Soc. **81**, 803 (1959).
[2] IR. W. S. Richardson and J. H. Goldstein, J. Chem. Phys. **18**, 1314 (1960).
[3] IR. G. R. Hund and M. K. Wilson, J. Chem. Phys. **34**, 1301 (1961).

FIGURE 10

Typographic Output. Photograph of a page from reference [32]. Copy was prepared using the GPSDC system and then translated to the code system of a photo composition machine.

TABLE 2. *Reactions of e_{aq}^- with water and transients from water*

No.	Solute and Reaction	pH	k(dm^3 mol^{-1}s^{-1})	Method	Comments	Ref.
1.1	H_2O $e_{aq}^- + H_2O \rightarrow H + OH^-$	8.3–9.0	$(1.6 \pm 0.1) \times 10^1$	p.r.	computer anal.; contains 7×10^{-4} *M* H_2.	Hart..66–0015
		8.3		p.r.	k detd. at 5–81°C to give $E_a = 4.5 \pm 1$ kcal mol^{-1}.	Fiel.67–0532
		11	$(2.2 \pm 0.6) \times 10^1$	p.r.	contains $Ba(OH)_2$ and 4×10^{-3} *M* formate ion; extrapolated to formate concn. = 0.	Swal68–0418
		> 7	2.7×10^1 (rel.)	γ–r.	c.k., assume $k(e_{aq}^- + NO_3^-) = 1.1 \times 10^{10}$, soln. contains 3×10^{-5} *M* $NaNO_3$ and 5×10^{-2} *M* glucose; pressures up to 8.85 kbar.	Hent.70–0056
1.2	D_2O $e_d^- + D_2O \rightarrow D + OD^-$	9.39	1.25 ± 0.5	p.r.	computer anal., D_2O soln. satd. with D_2.	Hart.68–0025
1.3	e_{aq}^-	–	$(6.5 \pm 1.0) \times 10^9$	p.r.	——	Dorf.63–0045
	$e_{aq}^- + e_{aq}^- \rightarrow H_2 + 2OH^-$	13	5×10^9	p.r.	——	Gord....63 -0050
		10.9	$(4.3 \pm 0.8) \times 10^9$	p.r.	——	Gord....63–0073
		13.3	$(5.5 \pm 0.7) \times 10^9$	p.r.	soln. in equil. with 100 atm. H_2.	Math.65–0009
		12	$(6.3 \pm 1) \times 10^9$	γ–r.	steady–state method, soln. H_2–satd., method leas reliable, k detd. at 10 – 93°C to give $E_a = 5.2 \pm 0.3$ kcal mol^{-1}.	Gott.67–0109
		11	6×10^9	f.phot.	soln. H_2–satd.	Schm.68–7143
		12.7	5.0×10^9 (cor.)	p.r.	apparent change in k with pH has been obs.	Brus70–0749
1.4	e_d^- $e_d^- + e_d^- \rightarrow D_2 + 2OD^-$	13.4	6.0×10^9	p.r.	computer anal., D_2O soln. contains 5.7×10^{-3} *M* D_2.	Hart.68–0025
1.5	H	10.9	$\sim 3 \times 10^{10}$	p.r.	——	Gord....63–0073
	$e_{aq}^- + H \rightarrow H_2 + OH^-$	10.5	$(2.5 \pm 0.6) \times 10^{10}$	p.r.	soln. is in equil. with 100 atm. H_2.	Math.65–0009
1.6	D $e_d^- + D \rightarrow D_2 + OD^-$	9.39	$(2.8 \pm 0.2) \times 10^{10}$	p.r.	soln. contains 4.5×10^{-3} *M* D_2 in D_2O.	Hart.68–0025
1.7	OH	10.5	$(3.0 \pm 0.7) \times 10^{10}$	p.r.	soln. contains only NaOH.	Math.65–0009
	$e_{aq}^- + OH \rightarrow OH^-$	11	3×10^{10}	p.r.	——	Gord....63–00730
1.8	OD $e_d^- + OD \rightarrow OD^-$	11.15	$(2.8 \pm 0.2) \times 10^{10}$	p.r.	computer anal., D_2O soln. of NaOD.	Hart.68–0025
1.9	O^- $e_{aq}^- + O^- \rightarrow 2\,OH^-$	13	$(2.2 \pm 0.6) \times 10^{10}$	p.r.	soln. in equil. with 50 atm. H_2, contains NaOH; not very reliable value.	Math.65–0009
1.10	O_2^- $e_{aq}^- + O_2^- \rightarrow O_2^{2-}$	11.1	1.3×10^{10}	p.r.	d.k. at 650 nm(e_{aq}^-); computer anal.	Grue...71–0171

FIGURE 11

Tabular data. Typeset material from reference [30]. Original copy was prepared on a punched paper tape typewriter in essentially the same format using half-line spacing for superscripts and subscripts and underlining to indicate italics. Changes in type size and font, use of inferiors and superiors and of rules were introduced by editing the GPSDC copy.

applications presented by users of the GPSDC system have had a major impact on design of modules. Each application has revealed desires (or demands) that reflect the user's wish to invoke the capabilities of the total publishing process, human and machine, as opposed to printing. The users have agreed to a general approach. They know that their next job will require a variant. They also have learned that special programming on their part will be minimized.

The modular structure of the system is based on a clear separation between devices, which may use any code formalism, and archival representation in GPSDC. As soon as possible in the work flow, the code stream from an input device is converted to GPSDC. All processing is done on the GPSDC form. As late as possible the records are converted for use on a specific output device. Another design criterion is that individual devices should be treated as members of a class the best of which is slightly more powerful than any known members.

One result of this general approach is that all of the input from typewriters is interpreted by one program, although six different types of machines have been used. Also, records written originally in all capitals have been "marked up" by textual substitution routines and then passed through the same program.

All editing is done with one program package that operates on GPSDC records. Interestingly enough, several parts of this package are simply modifications of programs built by others to operate on ordinary binary coded decimal (BCD) records. The same is true for reformatting and information retrieval.

6.6 General remarks. It has become clear to us from this experience that the expanded code used in GPSDC presents no bar to the development of a very extensive and flexible text-handling system. Input can be accepted from a wide variety of sources. Very different material can be edited and reformatted by common programs. Any output device appears to be accessible. It is also clear that an automated text processing system can be written in a high level language, and be written by many hands.

It is this experience that makes us confident that this or a similar system, based on standard codes, can be used profitably by data centers, both for their internal operations and in cooperation with each other.

Our experience also causes us to suggest that formal and informal associations of data compilers and users should increase the level of their interaction with the groups formally charged

with the task of developing standards for automated data processing. We are particularly sensitive to the need for development of flexible character string and character array processing facilities in machine independent languages at the FØRTRAN and CØBØL level. It seems to us that the broad adoption of a basic standard man-machine alphabet should make it possible for the developers of machine independent languages to assume a standard man-machine alphabet for data representation. However, in making this suggestion we must emphasize that at present GPSDC has no formal standing as a USA Federal, USA National or International standard. As of today GPSDC is part of an experiment. Its present users are aware that much of the GPSDC system is based on proposed, not formally adopted, standards. We need more experience with broader classes of users, particularly in connection with diagrams, before we will be completely confident about the general validity of many of our ideas. In addition, we must reemphasize that this paper was restricted, primarily, to discussions of the extension of the graphic character repertory. A major task ahead is the specification of additional controls and the development of revised prescriptions for recording data files on magnetic tape intended for interchange and dissemination as publications.

If GPSDC or any alternatives are to become formally recognized automated data processing standards, a significantly large body of users must exist and the prescriptions of their standards must be made known through proper channels. A number of potential channels exist.

Within the National Bureau of Standards the Institute for Computer Sciences and Technology has the formal responsibility for leading the development of standards for automated data processing. As a partial discharge of this responsibility it maintains an index [34] which sets forth an outline of the ISØ, ANSI and USA Federal efforts in developing standards for automated data processing. This index provides a good starting point for people interested in learning more about the formal processes of standards development.

Acknowledgements

The work reported here has been a continuing effort at synthesis under the restraints of developing standards. Some very important sources of certain of the ideas and concepts included in the synthesis have not been cited directly. It is a pleasure to acknowledge the benefits we derived from our study of the work of Feldman [35], Klerer [36], Mullen [37]

and Anzelmo [38]. We are particularly indebted to the late James R. Mullen for helpful criticism of our earlier work. During the later phases of the work one of us (DG) was able to examine a very important development at National Diet Library (Japan) [39] that provides a keyboard device for up to 5000 primitive characters spread over 14 fonts plus a number of complex symbols created using overstrike techniques. The system includes provisions for typesetting. This development demonstrates even better than we can that much larger character sets than those presented here can be handled successfully in an automated environment when the needs of the user are compelling.

In the actual implementation of hardware involving the design of type characters the assistance and advice provided by Carl Lucas of IBM Endicott and Paul Simms of Teletype Corp. was indispensible.

A number of colleagues and former colleagues at the National Bureau of Standards have made important contributions. Mrs. C. Messina and Messers. W. Evans, C. Albright, R. Chandler, R. McClenon, I. Soroka and R. Thompson have added special programs and assisted in a number of phases on systems maintenance. Mr. A. Weissberg's support was essential particularly in the early phases of the work and he made significant technical contributions to the design of the Taxywriter [6]. Mr. J. Hilsenrath has encouraged the use of the system and made many of the special applications possible.

A special note of appreciation is due Mrs. C. Seymour and Miss E. Hoffman who used the GPSDC system to prepare the typescript of this publication. The continuous participation of typists as experimenters has proved to be one of the most useful features of our program.

7. References

[1] International Organization for Standardization (ISO), Recommendation R 31

"Part II: Quantities and units of periodic and related phenomena" (Feb. 1958)

"Part III: Quantities and units of mechanics" (Dec. 1960)

"Part IV: Quantities and units of heat" (Dec. 1960)

"Part V: Quantities and units of electricity and magnetism" (Nov. 1965)

"Part VII: Quantities and units of acoustics (Nov 1965)

"Part XI: Mathematical signs and symbols for use in physical sciences and technology" (Feb. 1961) American National Standards Institute, 1430 Broadway, New York, New York.*

[2] International Organization for Standardization, "Rules for the use of units of the International System of Units and a selection of the decimal multiples and sub-multiples American National Standards Institute, New York, New York.*

[3] International Union of Pure and Applied Physics, Commission on Symbols, Units and Nomenclature, "Symbols, units and nomenclature in physics" Document ULP 11 (SUN 65-3) 1965.

[4] International Union of Pure and Applied Chemistry, Commission on Symbols, Terminology and Units, M. L. McGlashan, chairman, "Manual of Symbols and Terminology for Physicochemical Quantities and Units", Pure and Applied Chemistry 21, 3 (1971).

[5] "Proposed Revised American Standard Code for Information Interchange", Communications of the ACM 8, 207 (1965).

[6] "Modified tape-recording typewriter", Nat. Bur. Stand. Tech. News Bull. 50 No. 7, 118 (July 1966).

[7] "The Document Image Code". Nat. Bur. Stand. Tech. News Bull. 52 No. 4, 86 (April 1968).

[8] "Prototype General Purpose Scientific Document Writer Installed", Nat. Bur. Stand. Tech. News Bull. 54 No. 2, 35 (Feb. 1970).

[9] University of Chicago Press, A Manual of style, 12th ed. rev. 1969 Chicago, Ill.

[10] American Institute of Physics, Style Manual, 2d ed. 1959, rev. 1969 New York, New York.

[11] American Chemical Society, Handbook for Authors, 1st ed. 1967 Washington, D. C.

[12] ISO Technical Committee 97, Subcommittee 2 on Character Sets and Coding, "Proposed Revision of ISO Recommendation R 646, 7-bit Coded Character Set for Information Interchange ISO TC 97/SC 2 Document 566 (1971), American National Standards Institute, New York, New York.*

[13] ISO Technical Committee 97 Subcommittee 2 on Character Sets and Coding, "Third Draft ISO Proposal for Code Extension Techniques for Use with 7-bit Coded Character Set of ISO/646", ISO TC 97/SC 2 Document 562 (1971), American National Standards Institute, New York, New York.*

[14] American National Standards Institute, "Code for Information Interchange", ANSI Standard X3.4-1968, New York, New York.*

[15] Japanese Industrial Standard Committee, "Japanese Industrial Standard Code for Information Interchange" JIS C 6220-1969, Japanese Standards Association, Tokyo.

[16] USSR Committee on Standards, Measures and Measuring Equipment of the Council of Ministers, USSR, "Computers and data transmission equipment alphanumeric codes", GOST 13052-67 Moscow 1968.

[17] American National Standards Institute, "American National Standard for Bibliographic Information Interchange on Magnetic Tape", ANSI Standard Z39.2-1971 New York, New York.

[18] Teletype Corporation, "Model 37/300 Automatic Send-Receive Terminals", Teletype Model 37, Product catalog 37 CAT 5M469, Skokie, Illinois 60076.

[19] American National Standard Institute, "American National Standard Hollerith Punched Card Code" ANSI Standard X3.26-1970 New York, New York.

[20] Clamons, E.H., " Character Codes: Who Needs Them?", Honeywell Computer Journal 5, 143 (1971).

[21] Gottardi, R., "A Modified Dot-Bond Structural Formula Font with Improved Stereo-chemical Notation Abilities" J. Chem. Documentation 10 75 (1970).

[22] Nat. Bur. Stand. (U.S.), "Implementation of the Code for Information Interchange and Related Media Standards", Fed. Info. Process. Stand. Publ. (FIPS PUB.) No. 7 (1969).

[23] International Standards Organization, "Procedure for the Registration of Escape Sequences" ISO Draft Recommendation 2375, American National Standards Institute, New York, New York.*

[24] Hershey, A. V.,Preparation of Reports with the Fortran Typographic System, Tech Note TN-K/27-70, US Naval Weapons Laboratoratory, Dahlgren, Va. (1970).

[25] A.H. Philips, Computer Peripherals and Typesetting, Her Majesty's Stationary Office, London, 1968.

[26] British Standards Institution, "Specification for Metric Typographic Measurement", B.S. 4786:1972, London.

[27] Bulletin of Thermodynamics and Thermochemistry No. 12 (1969), No. 13 (1970), No. 14 (1971), No. 15 (1972), No. 16 (1973), University of Michigan Publication Distribution Service, Ann Arbor, Michigan

[28] McClenon, R., Evans, W. H., Garvin, D. and Duncan, B. C., "Description of the Magnetic Tape Version of the Bulletin of Thermodynamics and Thermochemistry, No. 14 (1971)" Nat. Bur. Stand. (U.S.) Tech. Note 760 (1973).

Idem. "Magnetic Tape Version of the Bulletin of Thermodynamics and Thermochemistry No. 14 (1971)", Nat. Bur. Stand. (U.S.) Magnetic Tape 4 (1973) (available from Nat. Tech. Information Service as COM-73-10627).

[29] Shimanouchi, T., "Tables of Molecular Vibration Frequencies Consolidated Volume I", Nat. Stand. Ref. Data Ser. - Nat. Bur. Stand. 39 (1972).

[30] Anbar, M., Bambenek, M. and Ross, A. B., "Selected Specific Rates of Reactions of Transients from Water in Aqueous Solutions", Nat. Stand. Ref. Data Ser. - Nat. Bur. Stand. 43 (1973).

[31] Thompson, R. C., Messina, C. G. and Hilsenrath, J., "Case Studies in Source-Automation of Technical Publications at the National Bureau of Standards", unpublished NBS working paper.

Hilsenrath, J., Messina, C. G., and Shimanouchi, T., "Source Automation of Tables of Molecular Vibration Frequencies", unpublished NBS working paper.

Messina, C. G., "SETDIC-ALGRID", unpublished FORTRAN programs for converting GPSDC to "Linotron" driver records.

[32] Donnay, J.D.H. and Ondik, H. M., general editors, Crystal Data Determinative Tables 3rd edition. Nat. Bur. Stand. and Joint Committee on Powder Diffraction Standards, Vol. I (1972), Vol. II (1973).

[33] Nat. Stand. Ref. Data Series - Nat. Bur. Stand. (U. S. Government Printing Office, Washington).
The volumes converted were NSRDS-NBS 9, 20, 26 and 31.

[34] Nat. Bur. Stand. (U.S.), "Federal Information Processing Standards Index", Fed. Info. Process. Stand. Publ. (FIPS PUB.) No. 12 (1971).

[35] Feldman, A., Holland, D.B., Jacobus, D. P., "Survey of Chemical Notation Systems", J. Chem. Documentation 3, 188 (1963).

[36] Klerer, M. and May, J., "An Experiment in a User-Oriented Computer System", Communications of the ACM 7, 5 290 (1964).

[37] Mullen, J.M., "Atom Typewriter Input for Computerized Storage and Retrieval of Chemical Structures, J. Chem. Documentation 7, 88 (1967).

[38] Anzelmo, F.D., "A Data-Storage Format for Information System Files", IEEE Trans. on Computers C-20, 39 (1971).

[39] "Kanji and Computer" Nat. Diet Library Newsletter No. 32, June 1971.

*ISO Documents, are usually available from the standardization organizations in member nations.

U. S. GOVERNMENT PRINTING OFFICE: 1974--542-653/132

NBS-114A (REV. 7-73)

U.S. DEPT. OF COMM. BIBLIOGRAPHIC DATA SHEET	1. PUBLICATION OR REPORT NO. NBS TN-820	2. Gov't Accession No.	3. Recipient's Accession No.

4. TITLE AND SUBTITLE

Complete Clear Text Representation of Scientific Documents in Machine-Readable Form

5. Publication Date: February 1974

6. Performing Organization Code

7. AUTHOR(S): B. C. Duncan and D. Garvin

8. Performing Organ. Report No.

9. PERFORMING ORGANIZATION NAME AND ADDRESS

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

10. Project/Task/Work Unit No.: 6307591

11. Contract/Grant No.

12. Sponsoring Organization Name and Complete Address (*Street, City, State, ZIP*)

Same as No. 9.

13. Type of Report & Period Covered: Final

14. Sponsoring Agency Code

15. SUPPLEMENTARY NOTES

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

Science and technology use a large variety of symbols to represent physical properties, chemical formulas and mathematical expressions.

Data centers that codify and evaluate physical properties need to use this conventional symbolism in their work. It is recommended that these data centers adopt the symbols and terminology specified by the various International Unions both in manual operations and in the creation of machine-readable data bases.

It is demonstrated that these conventional symbols can be produced by modern communications devices that are compatible with the international standard codes for information interchange. A set of characters suitable for representing scientific data and text is presented and proposed as an extension of the ISO information interchange code.

The use of this extended character code by computer oriented data centers at the National Bureau of Standards is described. The equipment needed for this level of performance and criteria for their selection are outlined.

17. KEY WORDS (*six to twelve entries; alphabetical order; capitalize only the first letter of the first key word unless a proper name; separated by semicolons*) Graphic character sets; information analysis centers; information interchange codes; recording typewriters; scientific computer technology.

18. AVAILABILITY

[X] Unlimited

[ ] For Official Distribution. Do Not Release to NTIS

[ ] Order From Sup. of Doc., U.S. Government Printing Office Washington, D.C. 20402, SD Cat. No. C13

[ ] Order From National Technical Information Service (NTIS) Springfield, Virginia 22151

19. SECURITY CLASS (THIS REPORT): UNCLASSIFIED

21. NO. OF PAGES: 55

20. SECURITY CLASS (THIS PAGE): UNCLASSIFIED

22. Price: 90 cents

NBS TECHNICAL PUBLICATIONS

PERIODICALS

JOURNAL OF RESEARCH reports National Bureau of Standards research and development in physics, mathematics, and chemistry. Comprehensive scientific papers give complete details of the work, including laboratory data, experimental procedures, and theoretical and mathematical analyses. Illustrated with photographs, drawings, and charts. Includes listings of other NBS papers as issued.

Published in two sections, available separately:

- **Physics and Chemistry (Section A)**

Papers of interest primarily to scientists working in these fields. This section covers a broad range of physical and chemical research, with major emphasis on standards of physical measurement, fundamental constants, and properties of matter. Issued six times a year. Annual subscription: Domestic, $17.00; Foreign, $21.25.

- **Mathematical Sciences (Section B)**

Studies and compilations designed mainly for the mathematician and theoretical physicist. Topics in mathematical statistics, theory of experiment design, numerical analysis, theoretical physics and chemistry, logical design and programming of computers and computer systems. Short numerical tables. Issued quarterly. Annual subscription: Domestic, $9.00; Foreign, $11.25.

DIMENSIONS, NBS

The best single source of information concerning the Bureau's measurement, research, developmental, cooperative, and publication activities, this monthly publication is designed for the layman and also for the industry-oriented individual whose daily work involves intimate contact with science and technology —*for engineers, chemists, physicists, research managers, product-development managers, and company executives.* Annual subscription: Domestic, $6.50; Foreign, $8.25.

NONPERIODICALS

Applied Mathematics Series. Mathematical tables, manuals, and studies.

Building Science Series. Research results, test methods, and performance criteria of building materials, components, systems, and structures.

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

Special Publications. Proceedings of NBS conferences, bibliographies, annual reports, wall charts, pamphlets, etc.

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

National Standard Reference Data Series. NSRDS provides quantitative data on the physical and chemical properties of materials, compiled from the world's literature and critically evaluated.

Product Standards. Provide requirements for sizes, types, quality, and methods for testing various industrial products. These standards are developed cooperatively with interested Government and industry groups and provide the basis for common understanding of product characteristics for both buyers and sellers. Their use is voluntary.

Technical Notes. This series consists of communications and reports (covering both other-agency and NBS-sponsored work) of limited or transitory interest.

Federal Information Processing Standards Publications. This series is the official publication within the Federal Government for information on standards adopted and promulgated under the Public Law 89–306, and Bureau of the Budget Circular A–86 entitled, Standardization of Data Elements and Codes in Data Systems.

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.

BIBLIOGRAPHIC SUBSCRIPTION SERVICES

The following current-awareness and literature-survey bibliographies are issued periodically by the Bureau:

Cryogenic Data Center Current Awareness Service (Publications and Reports of Interest in Cryogenics). A literature survey issued weekly. Annual subscription: Domestic, $20.00; foreign, $25.00.

Liquefied Natural Gas. A literature survey issued quarterly. Annual subscription: $20.00.

Superconducting Devices and Materials. A literature survey issued quarterly. Annual subscription: $20.00. Send subscription orders and remittances for the preceding bibliographic services to the U.S. Department of Commerce, National Technical Information Service, Springfield, Va. 22151.

Electromagnetic Metrology Current Awareness Service (Abstracts of Selected Articles on Measurement Techniques and Standards of Electromagnetic Quantities from D-C to Millimeter-Wave Frequencies). Issued monthly. Annual subscription: $100.00 (Special rates for multi subscriptions). Send subscription order and remittance to the Electromagnetic Metrology Information Center, Electromagnetics Division, National Bureau of Standards, Boulder, Colo. 80302.

Order NBS publications (except Bibliographic Subscription Services) from: Superintendent of Documents, Government Printing Office, Washington, D.C. 20402.