A Standard Code for Special Typographic Character Identification

Stanley Rice

An industry-wide standard code to identify typographic characters and their uses for electronic character generation is proposed. The code would facilitate both traditional and automatic analyses of character sets and provide a mutually intelligible communication channel for author, editor, designer, and compositor. Reactions to the code are solicited.

It is considered highly desirable that there exist an industry-wide standard code to identify typographic characters and their uses—especially those that are not on standard minimum keyboards and for which no standard machine-readable code now exists.

This code would facilitate both traditional and automatic analyses (by any system) of character sets required by any manuscript, and would provide for the first time a mutually intelligible and hardware-independent communication channel for author, editor, designer, and compositor.

New methods of character generation make such a code highly desirable in the very near future. It should greatly facilitate the flow of information, especially in scientific material, and for complex material generally.

A useful code would represent each possible character by a simple but faceted structure capable of representing, at least, class and character identification, and typographical use features such as mode, weight, condensation or expansion, relative size, and relative vertical position in a line.

A preliminary description of a general code constructed along these lines is provided here. Your reactions are solicited and may be sent to The American Institute of Graphic Arts, 1059 Third Avenue, New York, N.Y. 10021.

A Preliminary Description

There is currently no common language or code for typographic specification of extensive character sets—let alone one that can meet current needs associated with newer typesetting and type-generating methods. Such inadequate specification systems as now exist are tied to traditional hardware systems and are usually associated with the particular needs and growth patterns of those systems rather than with the needs of author, editor, designer, and the individual typesetter.

Any graphic image that will be useful as a typographic character should be specifiable in simple and unambiguous terms, both as to its identity and as to generally defined potential typographic forms and characteristics. This specification should be possible in written notation and on all common typewriter or machine-readable keyboards—for human inspection, machine-readable coding in any system, and for optical character recognition by any system.

A simple code using only decimal digits and alphabetic caps can make these aims practical. All characters then become intelligible and machine-readable in any system and for any part of the communication chain. Any keyboard (or handwriting) will do to write the code, because there need be no ambiguities in any system between the characters used.

Subject to further study and criticism, we feel that there are certain essential character identification facets basic to typographic character discrimination. But there are others, often currently used, that are not essential to a general code.

A code is called for that will be based upon the practical needs of the entire communication chain, to be expressed in descriptive terms appropriate to those general needs. It must obviously, for example, be subject to both manual and machine interpretation. Surely it is obvious that many private code systems, such as are now springing up, will be much against the interests of the entire communications industry.

Also, it should be obvious that the assignment of ad hoc character substitution schedules, as in current practice, in no way obviates the growing necessity to define the actual identification and use facets of the special characters themselves—in specific and quantitative ways.

A general code will define types of characters, not individual character designs. Individual designs will be identified by their membership in a carefully defined set described by the code. When a general code is agreed upon, letter generating agencies of all kinds can then provide their own correlated listings, showing actual designs for characters available for sale within the various classifications described.

Some of the non-essential facets that are omitted from the general classification code proposed are: natural language character names (e.g., "pi"), typeface names (specific designs such as Baskerville), point size designations, set or unit values, font information (the collections accessible at one time on specific hardware), mixing or other mechanical limitations, codes for hardware access, and bearing or side spacing. These omitted facets are important only for specific hardware and specific character design, not for a general character identification code.

The actual structure of the descriptive code proposed is as follows.

General Code for Typographic Character Identification

Character identification

0 Class identification (see separate schedule), alphabetic
1–3 Identification within class, numeric

Typographic use

4 Style and use identification (see separate schedule), alphabetic
5 Mode (see separate schedule), numeric
6 Weight (0–9 scale, see notes), numeric
7 Condensation-expansion (0–9 scale, see notes), numeric
8 Size or vertical coverage (0–9 scale, see notes), numeric
9 Vertical position (0–9 scale, see notes,) numeric

The characters on the most common typewriter keyboard (e.g., IBM #101 Correspondence Keyboard) will not be specified by the code, in common practice. Their identity will be represented by any common code in use. But since such codes provide no typographic use information, the code provides that they also may be modified by the general code to provide such information. Also, their identity

can be "identified" in the common code, if desirable, by their three digit decimal code form (e.g., EBCDIC or BCDIC).

Characters that are not on the common typewriter keyboard are represented by the first four digits of the code only, if current setting is assumed and only the character identification is needed. If typographic use information is also needed, the full identification code is used, as proposed above.

The purpose of the general code is to identify all typographic characters that now exist and that may exist in the future—according to the possible combinations of their most important functional facets. All characters will fall within the defined ranges, or such is the aim (see notes). And within the defined ranges the values of the important facets may vary by small amounts. Of course, the values of the non-essential facets (not covered by the code) may vary in infinitely many ways.

The actual descriptive capacity of the code in full use is about 67½ billion combinations of identification and the typographic use facets. And within these limits there are infinitely many actual designs possible. There are 25,974 character identifications and 2,600,000 facet combinations.

The assumption in all this is that since typographic output flexibility is becoming increasingly prevalent, the author and the publisher should specify the less common typographic characters in general terms. These specifications can then be interpreted in terms of specific designs available or to be made available. Authors and publishers should not be governed now by past hardware limitations, or by the simple lack of a language to make their needs clear.

0 *Class Identification* (Classes are mutually exclusive)

X Characters represented in standard codes. These may be characters in the codes but not on the standard typewriter keyboard, or those on the keyboard when it is desirable for some reason to identify them as part of the general code.
A Accents in Romance languages (floating or non-floating)
Z Accents in non-Romance languages (floating or non-floating)
M Mathematical and logical—operations
W Mathematical and logical—non-operations
S Scientific (not mathematical or logical)
P Phonetic and teaching alphabets
R Reference and punctuation
C Cyrillic and Greek
H Hebrew and Arabic
N German
F Musical
D Decorative or pictorial (stylized and general use only)
G General geometrical (excluding 4C, 4Q, 4R, 4T)
Q Character augments (floating "non-space" character supplements—circles, underlines, cancels, etc.)
K Combination characters—logotypes, diphthongs, mnemonics
L Ecclesiastical, fraternal, commercial, monetary
B Bars, brackets, braces, rules, leaders
O Outline form
V Reverse form
I Multi-line interpretation for position #8
U Unclassifiable by classifications above

Letters E, J, Y and the many unassigned sequences within other letters are available for future needs. Obviously the above is simply an assignment for purposes of illustration and should be the subject of study.

Each class identification letter is followed by a three-digit number identifying 999 possible characters in that class.

4 *Style and Use Identification*

U Unclassified for style, or only one style permitted
A Roman—taper serif (general category, "universal" or transitional)
M Roman—taper serif, distinctly modern
O Roman—taper serif, distinctly old style
I Roman—sans serif
H Roman—slab serif
S Script or cursive (pointed pen or brush derived)
E Calligraphic (edged pen derived)
P Printout style or typewriter style (equal letter widths)
B Black letter ("Old English," etc.)
C Circle (solid, e.g. bullet)
Q Square (solid)
R Star (five pointed, solid)
T Triangle (equilateral, solid)

5 *Mode*

0 Mode irrelevant to character in question
1 Roman caps
2 Roman lower case (if arabic numeral, non-lining)
3 Small caps
4 Italic caps, slant #1
5 Italic caps, slant #2
6 Italic caps, slant #3
7 Italic lower case, slant #1 (if arabic figure, non-lining)
8 Italic lower case, slant #2 (if arabic figure, non-lining)
9 Italic lower case, slant #3 (if arabic figure, non-lining)

These codes for mode are here used as part of the general identification code. There would obviously be an advantage in using them to denote mode in any current setting—to remain in effect until an end delimiter.

A test could determine whether the first character after the parens is alphabetic or numeric. If alphabetic the general code is indicated, if numeric the mode code above is indicated. If desirable, the mode code can be followed by the #6 (weight), the #7 (condensation-expansion), the #8 (size only), and the #9 position

(vertical position)—but only in this order and with no omissions as far as the positions are extended. See the pages that follow. End is signalled by delimiter.

Notes on the Facets of the General Code

0 *Class Identification*
All classes are considered to be mutually exclusive in this category. One of the few limits to the code as it is here described is that owing to this class mutual exclusion only one member of this "zero" set may be described in respect to one character. For example, this means that only typewriter keyboard characters and four general geometric shapes (4C, 4Q, 4R, 4T) may be specified as outline form, reverse, or multi-line characters. Outline characters cannot be directly specified to be reverse, and reversed or outlined characters cannot be designated as multi-line. The reverse field size can be defined by the eventual point size and set width.

If a character is here defined as multi-line, the number of vertical lines it is to occupy will be defined by "I", which then modifies #8 position to mean vertical coverage rather than size.

Floating accents and character augments are "non-spacing" characters that require another character, the one following the code, for completion.

1–3 *Identification within Each Class*
999 characters may be defined within each class. Unassigned letters and sequences are to be used for future necessities. Oriental characters are omitted because of sheer bulk but a reference system to other tables could be used. This schedule needs considerable study and constitutes the structure to which each character is assigned in the general code.

If simple identification in current setting is the only necessity, only the class identification and these three digits need be used.

4 *Style and Use Identification*
Only clear-cut cases of taper-serif characters should be designated "old style" or "modern." Most serifed design adaptations should be classed "general," which means "universal" with respect to serif-type styles. Likewise this classification is used to discriminate

among the "universal" characters that can be used with all the broad style groups named.

5 *Mode*

Most characters except common alphanumerics would be classified "0" because most of them have no "mode" (e.g., phonetic characters have no caps or italics). The ranges of predefined slant for the italics should be a continuous range, roughly "sloped roman," traditional italic, and extreme-slope italic.

Note that bold-face versions of any mode noted in this facet will be generated by specifications in #6, below. Likewise condensed or expanded versions are generated by specifications in #7, below; and relative sizes by #8, below.

As noted, the lower-case mode, if it is applied to an arabic numeral means "non-lining."

6 *Weight*

0-9 scale provides ten ranges of boldness. Scale range #4 should probably be "normal" (an arbitrary definition based on usage, one critical dimension being named). Minimums and maximums for the range would be defined in terms of the ten-point size.

7 *Condensation-Expansion*

0-9 scale provides ten degrees of condensation-expansion. Scale range #4 should probably be normal, an arbitrary definition based on usage. Maximum expansion could indicate double the width of the normal form and maximum condensation one half the width of the normal form.

Characters that should not be condensed or expanded would always be specified as #4.

8 *Size or Vertical Coverage*

0-9 scale provides ten ranges of size in tenths of one line of the setting size, exclusive of leading. This is the normal meaning, but if class I is designated under class identification, the meaning here is changed to vertical coverage expressed in numbers of lines of current setting size, top of cap to bottom of descender. Up to ten lines, but note that numbering starts at zero.

Both tenths of lines and lines up to ten are relative size. The

eventual or actual size of the generated character is a function of the point size specified—not a part of the general identification code.

9 *Vertical Position*
0–9 scale provides ten ranges for location of the mid-point of the character. Each range is one-tenth of one line. Relative position in the line is what is expressed. Actual position is a function of the specified size and design.

When a manuscript is rendered machine-readable, the general code, or a referenced substitute of one or two keyboard characters, is keyboarded in place of the necessary non-keyboard characters. Most manuscripts, obviously, have a rather limited subset of special characters that must be specially coded. In practice these characters would be looked up once in the standard code book and listed as they are encountered, by the person marking the manuscript. (The author may already have used the code, of course.) The codes may then be assigned any convenient and arbitrary table look-up reference, and could be marked and/or keyboarded somewhat as follows (if the table look-up references were fewer than 26 characters, making possible alphabetic substitutions):

For the logical expression $\phi \rightarrow \psi$ we can type, say: (c,p,y)

A typical full code might be X076U04428 . This might stand for a non-standard-keyboard character having a standard code decimal designation 076 (EBCDIC character "less than" <), with unclassified style (here, one style existing), irrelevant mode, normal weight, normal condensation-expansion, small size (range #2), in superor position (range #8).

Another typical code might be as follows, if only typographical use information is needed: M88621 perhaps modifying a standard typewriter character coded in a standard code of any type. The typographic use indicated is as follows: modern roman type style, lower-case italic of #2 range slope (if a numeral, non-lining), boldface (#8 range), slightly expanded letterform (#6 range), small size (#2 range), and in an inferior position (#1 range). Possibly an inferior to an inferior.

If simple identification for current setting is all that is needed, only the classification letter and the three digits for identification

within class are needed; e.g., the code P058 for the phonetic character ð.

Table look-ups for common sets of special characters used by certain compositors can be arranged. And in general, by means of this code all characters can be accurately and rather easily expressed and interpreted by all parts of the communication chain, from the author to the output device. All tapes can be rendered mutually intelligible and revisable. It makes no difference whether the machine-readable keyboarding is by paper or magnetic tape, or for optical character recognition by any system. The same string of universally available characters will produce character identification and/or typographic use information, for any character assigned a place in the general code.

The communications industry should no longer be governed by traditional hardware conventions, or by the simple lack of a language for making specifications clear. Only if the needs are clear can compositors interpret them in terms of hardware and software and provide adequate preplanning. Only then can the communications channel operate with a reasonable minimum of noise in respect to character identification.