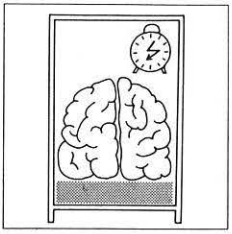
OCR-B: A Standardized Character for Optical Recognition

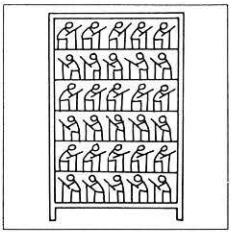
Adrian Frutiger
Collaborators: Nicole Delamarre and Andre Gürtler

OCR-B is a typefont especially developed as an international standard for optical recognition by electronic computers. It includes figures, upper- and lower-case letters, and certain related symbols. The background leading up to the development of OCR-B is discussed. Basically the problem was two-fold—to design a typefont (1) that could be automatically read by machines, and (2) that would be aesthetically accepted by the human eye. The design of OCR-B is examined in light of these requirements, and examples are shown.

The typefont presented here by Adrian Frutiger is the result of a standardization study to define a set of shapes which could be read accurately by electronic computers through optical reading equipment and which were not offensive to human taste. The definition of the typefont embodies compromises between the printing and reading requirements of many types of machines and the many typographical discriminations and habits of people in countries using a roman alphabet. Many opposing points of view were expressed within a design group which, as an engineer at Compagnie des Machines Bull, I brought together within the framework of the European Computer Manufacturers Association. Thanks to the enthusiastic collaboration of Adrian Frutiger, it was possible to reconcile his vast knowledge of typographic traditions and ruthless mathematical criteria.

The OCR-B font has been donated to public domain and now makes its own way along the careful paths laid down by ISO (International Organization for Standardization). Let us hope that it will soon replace many less decipherable predecessors on computer-made documents which have been part of our daily routine.—Gilbert Weill, Directeur de Programme au Centre National d'Etudes Spatiales, Paris

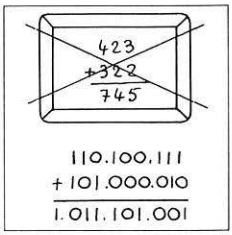




The Computer

Almost all human activities are now in some way related to the work of electronic computers. More and more, computers are the essential basis for the study of problems in scientific, economic, or sociological fields—even the problems of simple daily life.

Owing to their memory ability and their speed, these computers are able to perform the work of thousands of men. Within a few moments, hundreds of thousands of computing operations, comparisons, decisions, are made by the machine; whereas one hundred men equipped only with pencils would labor one hundred years to work out the same problems—and always with the possibility of error, taking into account the fallibility of the human brain.

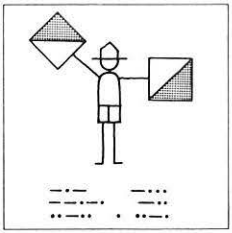


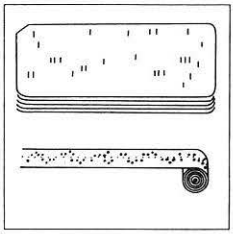
The Computer Cannot Read

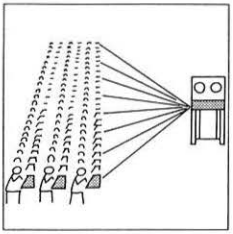
But the computer does not use the same language as men; it cannot work directly with figures and with our alphabet. It knows but two units: impulse and lack of impulse; it works with the binary system, i.e., every figure or letter is split up into two code signs.

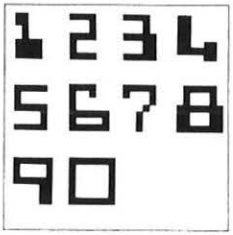
To represent data under coding forms is not new: the Morse code represents figures and letters by groups of dots and dashes, which are transmittable by two very simple movements.

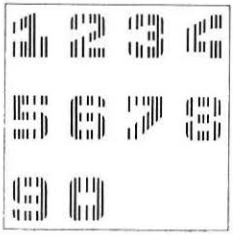
The computer communicates with our world in the same way a blind man reads in Braille: it scans a medium—punched tape or card, or magnetic tape—in order to find impulses in the form of holes or magnetic energy.

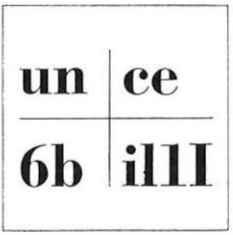












The Bottle-neck

Such tapes or cards are obtained by manual typing on a typewriter connected with a punching or magnetizing device. Every typed sign is converted into code. But the vast working abilities of the computer are inconsistent with the slowness of the human procedures necessary to change written data into code.

Many typists are then required to provide even a dropper-feeding of the computer; for within one hour a good typist can type only about 10,000 characters, when in the same time a computer is able to record ten million coded signs.

The Magnetic Characters

For many years electronics manufacturers, fully aware of this constraint, have been looking for a solution: automatic reading. Their research first led to the creation of specialized magnetic characters in which the code is directly incorporated into the image of the sign, that can then be read both by humans and by machines. Such systems have already led to automatization of banking procedures.

But using these characters demands that an excellent printing quality be maintained. Moreover, the distortions of our normal characters necessary for this system are almost unbearable to the human eye.

It Is Necessary to Have the Machine Read Our Writing

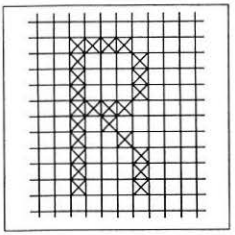
The setting up of reading devices able to identify automatically the configurations of our alphabet and to convert them into code becomes more and more essential. But we have to take into account the basic differences between the human and the machine reading processes; we do not read letter by letter, but by batches of syllables and words. Moreover, it is possible for our intelligence to deduce the meaning from the context. In fact, we quite naturally disregard similarities in the usual forms of letters. On the other hand, the computer reads only letter by letter and is unable to distinguish between such small differences of shapes as, for example, i, 1, l, and I.

Moreover, the quality of printing, which is poorest in typewritten material, is also to be taken into account; here is an enlargement of a typewritten figure 6, which shows important irregularities in lines.

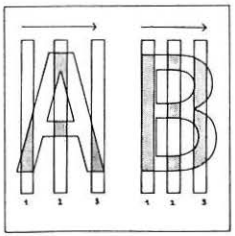
How

How is a computer able to differentiate the various shapes of letters, figures, and signs of our alphabet? Some illustrated explanations follow:

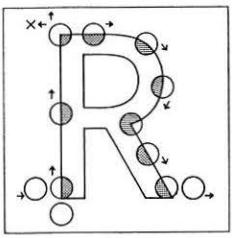
Key and Keyhole. Twenty-six keyholes corresponding to letter shapes are "programmed" in the reader. The letter which is read passes over the shapes, one by one, like a key searching for its corresponding hole. When it is inserted into the right opening, the code corresponding to the letter is found.



The Grid. A rectangle is divided into a certain number of small squares. All the letters must be inscribed inside this area. The shapes are conceived somewhat like the cross-stitch of our grandmothers' embroidery. In the reader, every small square is in fact a sensitive cell which records "dark" or "clear," that is, impulse or non-impulse. In this way the letter is coded.

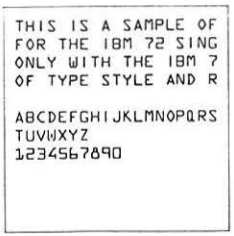


Automatic Reading System. In automatic reading the text is scanned by means of an elongated cell. If it recognizes, for example, that the left part of the letter about to be read is a complete stroke, it knows that this letter cannot be an A, but could be a B, not a C, but a D, or an E, or an F, etc. Scanning next the central part, then the right side, it is able to determine by process of elimination which letter is being read.



Another Automatic Reading System. Another way to identify the shape is to have its outlines explored by a cell programmed to search for the semi-darkness position—i.e., one half of the cell must always be clear, the other half dark. Thus, it automatically follows the outer edge of the letter and records the movements, which are then converted into code.

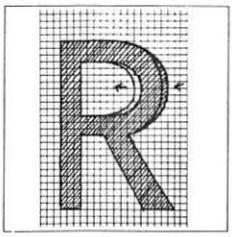
A Babel of Computer Languages Must Be Avoided
It is natural that every important computer manufacturer tends to develop his own optical reading method. But this leads to the danger that every company will create its own system of writing, corresponding to its reader requirements. For use at an international level confusion would be unavoidable, the machines of one manufacturer being unable to read the writings of the others.



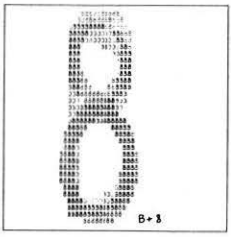
In the United States a special alphabet has been developed over the past several years in which the shapes of letters are strongly different from one another. The automatic reading of this alphabet is relatively easy, but such ease of recognition is to the detriment of the alphabet's appearance.

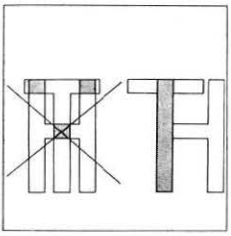
ECMA
The European Computer Manufacturers Association, conscious of the dangers of dispersal of methods, as well as of the neglect of the aesthetic and ethical aspect of the problem, has endeavored since 1963 to implement the full specifications for the creation of an optical reading alphabet including:
—a complete set of characters: figures, letters, signs, European accents;
—an optimum differentiation between every sign shape;
—an aesthetic appearance as close as possible to typographic characters.

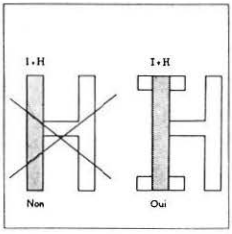
It was imperative to take into account *every* composition and printing process: fonts, typesetting and printing machines, typewriters, speed-printing machines, steel and plastic address plates, letterpress, offset, gravure, xerography, etc.

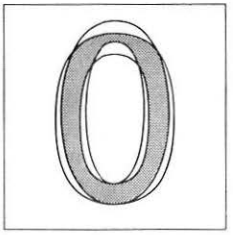


A Method
At the beginning the main problem was how to adapt a grid fine enough to be acceptable for all reading processes considered by every manufacturer and, at the same time, provide a screen fine enough to allow the design of harmonious shapes. On this ruled surface the development of the letterforms was carried out.









It was necessary to establish a method of measurement for comparison of shapes. The letter surface is punched in a card point by point. Then a computer compares every letter with all the others in the alphabet. It gives the results of the comparisons by superimposing the most critical groupings—for example, the 8 and the B shown here are easily discerned.

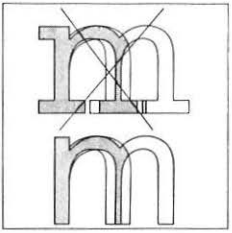
The comparison is carried out with a constant search for the most complete superimposition. For example, H and T are not compared with regard to their symmetrical positions but with regard to their largest common surfaces.

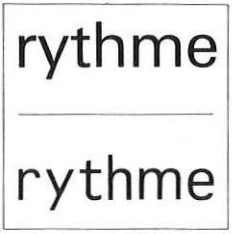
The Excessive Similarity of Our Letters

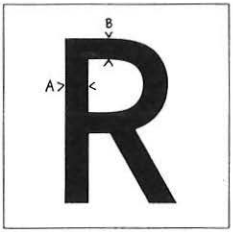
A character must never be fully included in another one; each must be differentiated by elements which are unique to it in comparison to the other one. For that reason, for example, it is imperative that the I should have serifs; in spite of opposition from the strictly typographic point of view, this is the only practical solution.

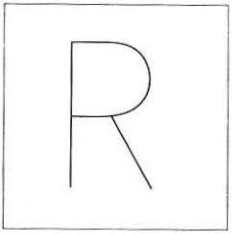
For the same reason, lower-case letters i, j, and l, capital I, figure 1, and the exclamation point have had to be more clearly differentiated than normally allowed by typographical use.

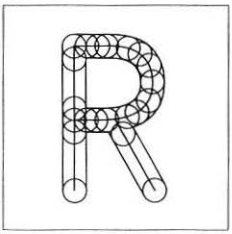
The greatest care must be exercised to distinguish clearly between certain figures and certain upper-case letters, chiefly between such symbols as 0 and 8 or B and S. In order to obtain that clear differentiation, it was decided from the beginning to decrease the height of upper-case letters in relation to figures and ascenders. This principle has also led to keeping all of the lower-case letters wide open and relatively extended, as well as simplifying placement of accents above upper-case letters.











Why Sans Serif?

All these considerations have from the beginning eliminated the choice of a type face with serifs. The repeated presence of horizontal strokes would have considerably increased the letter likeness, uselessly creating minor shapes common to all letters.

A relatively large space between letters is imperative in order to avoid any confusion when a cell is reading. In fact, basic principles of good typography have been violated to some extent. This example shows, above, a normal representation of a word in type; and below, the same word in OCR-B; spaces between r, y, and t have become quite large.

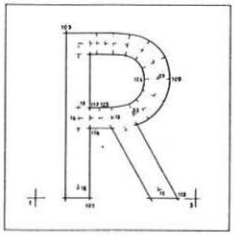
Stages of Elaboration

The first drawings were made in the traditional style. Curves are non-geometric and at the beginning are governed only by aesthetic laws. As far as possible, the drawing includes all the usual gradations of thickness between downstrokes (A) and upstrokes (B) (horizontal and vertical strokes), respecting the imposed thickness of minimum and maximum tolerances.

After testing final drawings of the traditional printers' shape, it was necessary to search for the centerline of all these non-geometric outlines. It is, in fact, this centerline which becomes the very basis, the skeleton, of all additional processes. It was, therefore, important to create first the printers' style, the most demanding from the aesthetic point of view, then to adapt it in a geometric way.

This design of the centerline automatically becomes the basic "blueprint" for typewriter font cutting. A drawing of an ideal enlargement of the typewriter font has been made around the skeleton lines (in a character called "sausage," with a constant stroke thickness).

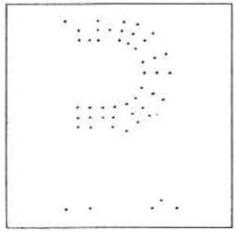
In order to meet the needs of various fields of application, the standardization of the alphabets in several sizes proved necessary. The enlargements were strictly proportional from the initial size, but with different multipliers for vertical and horizontal



6-39

CARACTERE R CHARACTER

N	Y1	X12	Y2
1	2000.	820.	2000.
2	1333.	2000.	1222.
3	2000.	3183.	2000.
101	2000.	1610.	2027.
102	2000.	1257.	2027.
103	4443.	1253.	4824.
104	4443.	1693.	4824.
105	4443.	1940.	4824.
106	4427.	2187.	4804.
107	4333.	2476.	4694.
108	4100.	2537.	4423.
109	3773.	2780.	4769.
110	3440.	2707.	3707.



measures. This "elongation" has been obtained as follows: along the outlines and centerlines of the printers' design, numbered measure points have been marked; every point refers to two measures: x = dimension measured from left vertical reference line; y = dimension measured from a horizontal base line.

A computer has calculated the new position for every point, and a pointing machine has set up reference points for the new size, from which it has been possible to establish the designs.

Trial Compositions

This composition specimen has been made in photocomposition, so that its appearance could be checked in actual composition. It is easily detected that widths of the different characters are variable. However, the exceptional width of the m, for which the maximal frame has been enlarged, is to be noticed; this letter is actually the most difficult one to condense without destroying the typographic appearance of the line.

Le principal goulot d'étranglement dans la suite des opérations de traitement réside dans l'obligation de préparer manuellement, à partir de documents conçus pour l'homme (états imprimés, fiches, registres) des documents directement exploitables par la machine, tels que des cartes perforées. Depuis de nombreuses années, les constructeurs de matériel électronique s'efforcent de

Here is also a first OCR-B cutting on the typewriter. Every letter is inscribed on the same width. Reading tests are to be made first of all with typed texts, representing a lower image quality than printed typography.

```
Les origines exactes de l'alphabet restent indistinctes.
Les caractères romains, qui sont à la base de notre al-
phabet actuel, s'apparentent aux caractères grecs lus,
à l'origine, alternativement de droite à gauche et de
gauche à droite.
```

The Reservoir

At present, testing on readers has been limited to the initial reservoir of forty-two characters, namely ten numerals, twenty-six upper-case letters, and six symbols. But a complete reservoir including a total of 111 characters has been designed in order to meet future needs. It would have been unthinkable to create an alphabet of upper-case letters without having determined the shape of a lower-case alphabet intended to go with it later on.

So that the typeface can be used in every country with a roman alphabet, nineteen characters and national signs have been added to the reservoir.

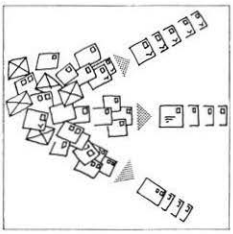
Standardization

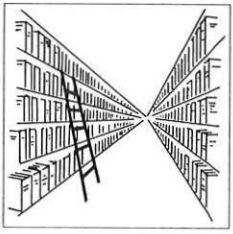
The whole of this work has been carried out with a view toward obtaining an international standardization for optical type. Under the name of OCR-B (Optical Character Recognition, class B) this new type has now become an international standard by which manufacturers are able to set up their reading machines.

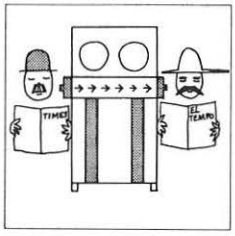
The Future

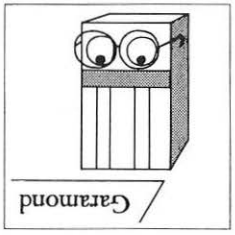
Automatic optical reading is likely to widen the bounds of the field of data processing. Let us cite as examples:

—Banks and insurance companies, which are already regular computer users, will have their task considerably simplified, owing to the standardization of one system for transcribing into code.









—Mail sorting, which poses more and more acute problems of time and manpower, can be done by automatic readers.
—Libraries can be consulted, and documentation, analysis, and selection made directly by the reading machine, without any limit of bulk, material, or time.
—Language processing and automatic translation can be carried out without manual transcription, whatever the geographic distance.

The creation of an optically legible type is an important step toward international co-operation in data processing. It may also be considered a success on a humane level; a new style of letters has not been created by the machine, but man tries to have the machine read the shapes which he created by long and difficult elaboration through centuries—from Egyptian stone-carving, through feather on parchment and the engraver's tool to the graphic artist, typographer, and printer in our time.

The future is still more encouraging in this respect: OCR-B will probably be but an intermediate stage. We can hope that one day "reading machines" will have reached perfection and will be able to distinguish without any error the symbols of our alphabets, in whatever style they may be written.