Designation: F149 – 92b (Reapproved 2009)

Standard Terminology Relating to Optical Character Recognition[1]

This standard is issued under the fixed designation F149; the number immediately following the designation indicates the year of original adoption or, in the case of revision, the year of last revision. A number in parentheses indicates the year of last reapproval. A superscript epsilon (ε) indicates an editorial change since the last revision or reapproval.

1. Scope

1.1 This set of definitions is intended for use by persons who, in the course of their duties, make use of OCR equipment or interact with operators of such equipment.

2. Referenced Documents

2.1 *ANSI Standards:*[2]

ANSI X3.17 Character Set for Optical Character Recognition (OCR-A)

ANSI X3.49 Character Set for Optical Character Recognition (OCR-B)

3. Terminology

3.1 *Definitions:*

adjacency—two OCR characters printed on the same line with character spacing reference lines separated by the proper space for the font and system.

alphameric—See **alphanumeric**.

alphanumeric—pertaining to a character set that contains letters, digits, and usually other characters such as punctuation marks. *Syn.***alphameric** .

alphanumeric character set—a character set that contains both letters and digits and may contain control characters, special characters, and the space character.

alphanumeric character subset—a character subset that contains both letters and digits and may contain control characters, special characters, and the space character.

average background reflectance—expressed as a percent, is the simple arithmetic average of the background reflection readings from at least five different points on a sheet.

average edge—an imaginary line bisecting the irregularities of the character edge.

backer printing—printing on the reverse side of the sheet. For OCR forms, the paper should have sufficient opacity so that printing on the back can't be seen on the front by the optical scanner.

background reflectance—a measurement of the brightness of paper referring to the amount of light reflected back from the paper at a particular point when that point is flooded with light, as compared with the known value representing absolute white (such as $BaSO_4$).

band—the light frequency spectrum between two defined limits; also light band.

banking—the alignment of the first graphic shape in a line with respect to the left (right) margin, by certain devices (that is, typewriters, line printers, etc.).

bar code—a binary coding system consisting of vertical marks or bars that, when read by an optical scanner, can be converted to machine language.

barium sulfate ($BaSO_4$)—a standard reflecting agent used to calibrate instruments for measuring the whiteness and reflectance of papers.

base line—a reference line used to specify the nominal relative vertical position of OCR characters printed on the same line.

basis weight—the weight in pounds of a ream cut to a specified basic size. The number of sheets in a ream is usually 500. The basic size for writing papers commonly used in OCR applications is 17 by 22 in. Also measured metrically in grams per square metre (g/m^2) and referred to as grammage.

blind ink—See **reflective ink**.

bridging—enlargement of a graphic shape beyond the COL, which produces undesired character fill in.

brightness—*in paper*, a characteristic of white paper measured in terms of reflectance in the blue and violet portions of the spectrum.

caliper—the thickness of a sheet of paper measured under specified conditions and usually expressed in thousandths of an inch (mils).

carbon paper—a sheet composed of a supporting substrate on one or both sides of which is a coating containing a

[1] This terminology is under the jurisdiction of ASTM Committee F05 on Business Imaging Products and is the direct responsibility of Subcommittee F05.01 on Nomenclature and Definitions.

Current edition approved Oct. 1, 2009. Published October 2009. Originally approved in 1972. Last previous edition approved in 2003 as F149 – 92b(2003). DOI: 10.1520/F0149-92BR09.

[2] Available from American National Standards Institute (ANSI), 25 W. 43rd St., 4th Floor, New York, NY 10036, http://www.ansi.org.

transferable (usually colored) material. The coating is of such nature that it will transfer in part or entirely to a copy sheet at the point of pressure contact.

centerline—the vertical axis around which character elements are located for letters, numerals, or symbols of an OCR font.

character—(*1*) a member of a set of elements upon which agreement has been reached and that is used for the organization, control, or representation of information. Characters may be letters, digits, punctuation marks, or other symbols, often represented in the form of a spatial arrangement of adjacent or connected strokes or in the form of other physical conditions in data media.

(*2*) a letter, digit, or other symbol that is used as part of the organization, control, or representation of data. A character is often in the form of a spatial arrangement of adjacent or connected strokes.

character alignment—the vertical or horizontal position of characters with respect to a given reference line.

character boundary—*in character recognition,* the largest rectangle with a side parallel to the document reference edge, each of whose sides is tangential to a given character outline.

character erase—an OCR graphic shape that will cover a single character or a single space and will be read by the interpreter as a deletion.

character outline limit (COL)—the minimum, nominal, and maximum limits of a given graphic shape.

character reader—an input unit that performs character recognition.

character reading—machine reading of alpha or numeric characters, or symbols, or both, by optical means (as opposed to optical mark reading).

character recognition—(*1*) The identification of characters by automatic means.

(*2*) See **magnetic ink character recognition; optical character recognition**.

character set—(*1*) a finite set of different characters upon which agreement has been reached and that is considered complete for some purpose, for example, each of the character sets contained in ANSI X3.17 and ANSI X3.49.

(*2*) an ordered set of unique representations called characters, for example, the 26 letters of the English alphabet, Boolean 0 and 1, the set of symbols in the Morse code, and the 128 ASCII characters.

character skew—the rotational deviation of the printed image from its intended orientation relative to a document reference edge.

character spacing—the pitch distance between adjacent characters.

character stroke width—the distance between the average edges of a character element.

character subset—a selection of characters from a character set, comprising all characters that have a specified common feature, for example, in each of the character sets contained in ANSI X3.17 and ANSI X3.49, the digits 0 to 9 may constitute a character subset.

clear area—that region of a document reserved for OCR characters and the required clear space around these characters.

*COL***—**See **character outline limit**.

contrast—(*1*) *in optical character recognition*, the difference between color or shading of the printed material on a document and the background on which it is printed.

(*2*) See **print contrast ratio**.

crowding—improper horizontal character spacing.

CVR—contrast variation ratio is the ratio between the maximum and minimum PCS within a graphic shape:

$$CVR = \frac{PCS,\max}{PCS,\min}$$

debossment—the depth of a print impression into the surface of a document.

dirt—*in paper,* refers to the presence of relatively nonreflective foreign particles embedded in the sheet. The size and lack of reflectance of the particles may be such that they will be mistaken for inked areas by an optical scanner.

document—a form designed as input to a document reader.

document reader—a scanning device that scans one to five lines of data in fixed locations on a document at a single pass. Generally, re-scanning of a portion of the document is not possible, one direction of the scan being provided by movement of the form past the reading head. The forms used generally don't exceed 8 to ¾ in. in width by 4 to ¼ in. in depth. Also see **page reader**.

drop out colors**—**See **reflective ink**.

drop out ink**—**See **reflective ink**.

edge irregularity—a variation in the stroke width of a printed character.

embossment—the height of raised print or raised surface on a document.

error—the substitution of one character for another.

error rate—the ratio of the number of character substitutions to the total number of characters read.

extraneous ink—any spot appearing within the "read" area, but outside the **COL** , caused by smear, tracking, or splatter that can be caused either in the manufacturing or while entering data on the form and can result in less optimum readability.

felt side—the top side of the paper in the paper manufacturing process as opposed to **wire side**. Optical scanning forms should be printed on the felt side.