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.

field—any group of characters defined as a unit of information.

field delimiter**—**See **field separator**.

field mark**—**See **field separator**.

field separator—a mark or symbol printed in scan ink that identifies fields to the scanner (Syn. *field mark*).

fluorescence—the property of emitting radiation in the visible range as a result of absorption of radiation in the ultraviolet range from some other source. Optical brighteners that have this property are sometimes added to paper to enhance its whiteness or brightness to the eye in normal lighting. The emitted radiation can cause erratic reflectance values.

flying spot scanning—*in optical character recognition,* a device employing a moving spot of light to scan a sample space, the intensity of the transmitted or reflected light being sensed by the photoelectric transducer.

font—a set of graphic shapes that may be alphabetic, numeric, or both and may include other symbols.

format—preprogrammed identification of fields to be read by an optical scanner.

free form (unformatted form)**—**a form on which the data appears in variable length fields. Preprinted symbols and guides are absent or minimal. Field delimiters are entered with the data.

grain long—paper grain direction in sheets of paper is parallel to the long dimension of the sheet.

grain short—paper grain direction in sheets of paper is parallel to the short dimension of the sheet.

group erase—an **OCR** graphic shape that will delete a group or string of three or more characters.

hand print boxes—restraints for controlling entry of scannable information by hand. Controlling of size, shape, and configuration of hand printed entries on an optical scanning form.

hand print character set—Refer to ANSI X3.45-182.

infinite pad method—*in optical character recognition,* a method of measuring reflectance of a paper stock such that doubling the number of backing sheets of the same stock will not change measured reflectance.

infrared response—a particular type of optical system used in some scanners. As a general rule, nonscan inks for this purpose are in the red portion of the color spectrum.

ink, OCR—Refer to ANSI X3.86-180.

interpreter—that part of the OCR system which analyzes the input data and determines what the individual characters are and what their relation is to each other.

ion deposition printer—a printer where ion charges are gated onto a dielectric drum. Toner is picked up by the charge, then transferred to the paper. Once the toner is deposited on the paper, it can be affixed by either pressure or heat fusion.

laser scanner—an optical scanning device that uses the intense monochromatic light beam given off by a laser as its source of illumination.

leading edge—the edge of a form that is used as a base for locating the first line of data to be scanned.

length/depth—the distance between the two edges of a form, reached by moving at right angle to a nominal data line.

light stability—*in optical character recognition,* the resistance to change of the color of the image when exposed to radiant energy.

line skew—the angular displacement of a line in relation to its intended position.

line spacing—the distance between the average base line of one line to the average base line of the next line.

machine language—a language designed for use by a machine, without translation.

magnetic ink character recognition, MICR—a recognition technology that utilizes ink capable of being magnetized and sensed. A practical application is E-13B, which is used primarily within the North American financial industry. E-13B consists of 14 characters printed to high specifications using ink with iron oxide pigments, or other inks utilizing ingredients capable of being magnetized.

magnetic printer—a printer in which magnetic signals are recorded onto a magnetic belt or drum. A magnetic toner is attracted to the drum and transferred to paper where it is fused to the sheet.

margin—the distance between any boundary of the printing area and the nearest parallel paper edge.

mark reading—machine reading, by optical means, of marks (usually vertical or horizontal bars) that have been manually entered.

mark scanning—the automatic optical sensing of marks usually recorded manually on a data medium.

mark sensing—machine reading of marks (usually pencil strokes) on a punched card, by using the conductive properties of the mark itself.

marking position—the area designated to mark information on a mark read form. Also called a **response position**.

mechanical disk scanner—a rotating scanning disk that breaks light reflection during the optical reading operation into a series of light points that are directed through the slit of a fixed aperture and onto the surface of a photomultiplier tube.

MICR—(*1*) An abbreviation commonly applied to the character set (E-13B) contained in ANSI X3.2-76 and X9.13-83.
(*2*) See **magnetic ink character recognition**.

millimicron (Mu)**—**a unit of length used in measuring light waves. The peak spectral response of a scanner is expressed in Mu.

moisture resistant paper—a category of optical scanning paper developed to meet unusual ambient or climatic conditions, for example, census forms or meter reading forms.

multifont reader—a reading device that can read forms containing intermixed characters printed in a number of fonts. Multifont reading eliminates the need to prebatch the input data by font prior to submission to the scanner.

multiple font reader—a reading device that can read more than one type font, but only one font may be read at a time.

MR-8—the original optical scanning test device that measures the amount of reflected light in millivolts.

noise—(*1*) random variations of one or more characteristics of any entity such as voltage, current, or data.

(*2*) a random signal of known statistical properties of amplitude, distribution, and spectral density.

(*3*) loosely, any disturbance tending to interfere with the normal operation of a device or system.

nonreflective ink—See **scan ink**.

nonread ink—See **reflective ink**.

nonscan ink—See **reflective ink**.

numeric—a machine **vocabulary** that includes only the primary numbers as contrasted to **alphanumeric**, which includes both letters and numerals.

OCR—See **optical character recognition**.

OCR-A—an abbreviation commonly applied to the character set contained in ANSI X3.17.

OCR-B—an abbreviation commonly applied to the character set contained in ANSI X3.49.

OMR (optical mark reading)—the process of identification of marks by an optical scanner.

off-line—pertaining to the operation of a functional unit not under the direct control of the computer.

on-line—pertaining to the operation of a functional unit when under the direct control of the computer.

opacity—the property of paper that minimizes the show-through of printing from the back side or the next sheet. The ratio of the paper reflectance with a black backing to the paper reflectance with a white backing.

optical brightener—a material often added to paper during its manufacture to improve its brightness or whiteness. These materials can cause erratic reflectance values when used with optical scanners that are sensitive to the short wavelength portions of the spectrum. See **fluorescence**.

optical character reader—an information processing device that accepts prepared forms and converts data from them to computer output media via **optical character recognition**.

optical character recognition (OCR)—character recognition that uses optical means to identify graphic characters.

optical scanner—(*1*) a scanner that uses light for examining patterns.

(*2*) a device that scans optically and usually generates an analog or digital signal.

page—a form on which many lines of data may be entered for reading by a page reader. Pages are larger in size than **documents**.

page reader—an electronic machine capable of reading full pages of printed data.

paper grain—the paper machine direction of paper.

paper grain direction—see **paper machine direction**.

paper machine direction—the direction of paper grain parallel with the direction of movement on the paper machine. It is also called *grain direction*. The direction at right angles to the paper machine direction is called the *cross-machine direction*, or simply, *cross direction*.

paper, OCR—paper used in **OCR** systems; refer to ANSI X3.62-79.

paper smoothness—the degree of irregularity of the surface of paper determined by the measurement of the flow of air between the paper surface and a plain surface under specified clamping and air pressure. The resistance to air flow increases as the paper goes from rough to smooth. With an instrument measuring the rate of flow, as the paper goes from rough to smooth, the number rating goes down. With an instrument measuring the time for a given volume to flow, as the paper goes from rough to smooth, the number rating increases.

paper weight—see **basis weight.**

PCS—See **print contrast signal.**

peaks—extraneous marks extending from the character outward past the COL.

pitch—the distance between one character reference point and the corresponding point on the next adjacent character.

porosity—*in paper,* the property that allows the passage of air through the sheet—an important factor in ink penetration, and also a quality that may affect paper feeding in some readers that have vacuum feeding mechanisms.

POS—an abbreviation for *p*oint *o*f *s* ale data entry systems where actual transactions are recorded by terminals operating on-line to a central computer. These systems frequently employ optical scanning as a means of capturing data.

print contrast ratio—*in optical character recognition,* the ratio obtained by subtracting the reflectance at an inspection area from the maximum reflectance found within a specified distance from that area, and dividing the result by that maximum reflectance. See **print contrast signal**.

print contrast signal—the relative value of the contrast of printing in relation to the paper background on which it is printed as defined by the following equation:

$$PCS_p = \frac{R_w - R_p}{R_w}$$

where:

R_w = maximum reflectance found within the area of interest to which the PCS of point p is referenced. (In measuring printed images, this area of interest should be a rectangle approximately twice the nominal character height by twice the nominal character width and centered on the character being measured), and

R_p = reflectance from a small measurement area centered on point p.

The reflectance R_w and R_p are measured within a circular area of 0.008 in. (0.2 mm) in diameter.

print quality—the interrelationship of printed material and imprinted material that affects the optimum performance of the scanner. Refer to OCR Print Quality Guideline, ANSI X3.99-83.

read area—one of several terms used to refer to the scan path or scan area.

read ink—See **scan ink**.

reference edge—the edge of the form used to align the form so that the nominal reading line will be parallel to the direction of scanning. Depending on the equipment used, this may be any edge of the form.

reflectance—the ratio of the response of a light sensor illuminated by diffuse reflection from the paper compared to that when the paper is replaced by a perfect diffuse reflector. A specially prepared surface of *barium sulfate* is considered to be a perfect diffuse reflector.

reflectance, absolute—the ratio of the total reflectance by a document to the total light incident on the document.

reflectance, diffuse—reflected light whose angle of reflection varies from the angle of incidence of the illuminating light, such as reflection from a rough surface.

reflectance, specular—reflected light whose angle of reflection is equal, or nearly equal, to the angle of incidence of the illuminating light, such as in reflectance from a mirror.

reflective ink—ink not sensed by the optical scanner, but visible to the human eye. Syn. *blind ink, drop-out colors, drop-out ink, nonread ink, nonscan ink.*

reject—a character located but not identified by an optical scanner.

reject rate—the number of rejects stated as a percentage of total items. Scanner items can be characters, marks, fields, documents, pages, etc.

repertoire—includes all of the characters and graphic shapes in an OCR imaging device system.

response position—the area designated to mark information on a mark read form.

scan—a search for information to be recognized by the recognition unit of the optical scanner, and the conversion of the optical signal to an electrical signal.

scan area—the area of a form that contains information to be scanned.

scan band—a strip across a document that passes directly beneath a scanning head of a reader.

scan ink—ink that is sensed by the optical scanner. Synonym for *nonreflective ink, read ink.*

scanner—(*1*) a device that examines a spatial pattern one part after another, and generates analog or digital signals corresponding to the pattern. Scanners are often used to mark reading, pattern recognition, or character recognition.

(*2*) See **flying spot scanner, optical scanner, mark scanning, optical scanner**.

serial printer—a printer where characters are printed one at a time.

skew—rotational deviation from correct horizontal and vertical orientation; may apply to a single character, line, or entire document.

source documents—the original materials or facsimiles from which input for a data processing system is derived.

special symbol/character—*in a character set*, a character that is neither a numeral, letter, or a blank, for example, virgule, asterisk, dollar sign, comma, period, etc.

spectral response—the variation in sensitivity of a device to light of different wavelengths.

spots—areas outside the maximum **COL**, which are contrasting with the background.

stacker—a device for accumulating processed documents in optical scanners and card readers.

stroke—*in character recognition,* a straight line or arc used as a segment of a graphic character. Each character is made up of a variable number of the strokes.

stroke average width—the average of actual stroke widths taken at points along the length of a stroke.

stroke centerline—*in character recognition*, a line midway between the two stroke edges.

stroke edge—*in character recognition,* the line of discontinuity between a side of a stroke and the background, obtained by averaging, over the length of the stroke, the irregularities resulting from the printing and detecting processes.

stroke width—*in character recognition,* the distance measured perpendicularly to the stroke centerline between the two stroke edges.

ticking—marks caused by the bottom of the upper case character while printing in the lower case, or opposite.

timing mark—a printed mark that controls the reading of a mark read field. The mark printed in scan ink tells the reader the location of information to be scanned.

throughput—the rate at which documents can be processed through an optical scanner. Usually expressed as "documents per minute."

transmitted light scanner—an optical scanner that operates by sensing light transmitted through paper instead of reflected from its surface.

turnaround document—a form produced by an electronic data processing system intended for future re-entry, possibly with added data, via an optical scanner.

ultraviolet response—a particular type of optics system used in some optical scanners. As a general rule, nonscan inks for this response will be in the violet portion of the color spectrum.

valleys—edge irregularities which result in an indentation in the side of a strike.

vertical field separators—(*1*) A vertical line separating data fields.

(*2*) See **field separator**.

visible response—a particular type of optical response system used in some scanners. There are very few reflective inks for this type of system.

vocabulary—See **repertoire** .

void—the absence of ink, or an area of significantly lower density of ink, within the confines of a character.

WAD—See **worst area difference.**

wand scanner—a hand-held optical scanner used in applications where it's impractical to transport data past a fixed read head.

white standard—a substance that reflects 100 % light and is used in calibrating test instruments. See **barium sulfate**.

wire side—*in paper,* the side of a sheet next to the wire in paper manufacturing.

worst area difference—one of the measures of how easily an interpreter can identify the characters of a font set, it is the range of the surface areas enclosed by the nominal COL's of the characters; the greater this range, the more easily the characters can be identified.