What is Non-Standard Encoding?

Let's start by looking at what **standard encoding** is and how it works.

Each English alphabetical and numerical character A, B, C or 1, 2, 3 and so on has a corresponding ASCII (American Standard Code for Information Interchange) Decimal or Hexadecimal code associated with it. Microsoft Word, Notepad and plain text editors use ASCII to display the desired characters.

ASCII code is the numerical representation of a character such as 'a' 'A' or '@' and can also be an action as shown in the chart below. ASCII-formatted text contains no font formatting information or font decoration such as bold or italic.

Example:
The string 'Solid Documents' in ASCII is equivalent to:
Decimal Value - 83 111 108 105 100 NULL 68 111 99 117 109 101 110 116 115
Hexadecimal value - 53 6F 6C 69 64 1 44 6F 63 75 6D 65 6E 74 73

Using the chart below, you can find the referenced values in decimal and hexadecimal for the example strings above. This table maps the value to the character that will display when these values are used.

Binary	Oct	Dec	Hex	Glyph
010 0000	040	32	20	space
010 0001	041	33	21	!
010 0010	042	34	22	"
010 0011	043	35	23	#
010 0100	044	36	24	$
010 0101	045	37	25	%
010 0110	046	38	26	&
010 0111	047	39	27	'
010 1000	050	40	28	(
010 1001	051	41	29	)
010 1010	052	42	2A	*
010 1011	053	43	2B	+
010 1100	054	44	2C	,
010 1101	055	45	2D	-
010 1110	056	46	2E	.
010 1111	057	47	2F	/
011 0000	060	48	30	0
011 0001	061	49	31	1
011 0010	062	50	32	2
011 0011	063	51	33	3
011 0100	064	52	34	4
011 0101	065	53	35	5
011 0110	066	54	36	6
011 0111	067	55	37	7
011 1000	070	56	38	8
011 1001	071	57	39	9
011 1010	072	58	3A	:
011 1011	073	59	3B	;
011 1100	074	60	3C	<

Binary	Oct	Dec	Hex	Glyph
100 0000	100	64	40	@
100 0001	101	65	41	A
100 0010	102	66	42	B
100 0011	103	67	43	C
100 0100	104	68	44	D
100 0101	105	69	45	E
100 0110	106	70	46	F
100 0111	107	71	47	G
100 1000	110	72	48	H
100 1001	111	73	49	I
100 1010	112	74	4A	J
100 1011	113	75	4B	K
100 1100	114	76	4C	L
100 1101	115	77	4D	M
100 1110	116	78	4E	N
100 1111	117	79	4F	O
101 0000	120	80	50	P
101 0001	121	81	51	Q
101 0010	122	82	52	R
101 0011	123	83	53	S
101 0100	124	84	54	T
101 0101	125	85	55	U
101 0110	126	86	56	V
101 0111	127	87	57	W
101 1000	130	88	58	X
101 1001	131	89	59	Y
101 1010	132	90	5A	Z
101 1011	133	91	5B	[
101 1100	134	92	5C	\

Binary	Oct	Dec	Hex	Glyph
110 0000	140	96	60	`
110 0001	141	97	61	a
110 0010	142	98	62	b
110 0011	143	99	63	c
110 0100	144	100	64	d
110 0101	145	101	65	e
110 0110	146	102	66	f
110 0111	147	103	67	g
110 1000	150	104	68	h
110 1001	151	105	69	i
110 1010	152	106	6A	j
110 1011	153	107	6B	k
110 1100	154	108	6C	l
110 1101	155	109	6D	m
110 1110	156	110	6E	n
110 1111	157	111	6F	o
111 0000	160	112	70	p
111 0001	161	113	71	q
111 0010	162	114	72	r
111 0011	163	115	73	s
111 0100	164	116	74	t
111 0101	165	117	75	u
111 0110	166	118	76	v
111 0111	167	119	77	w
111 1000	170	120	78	x
111 1001	171	121	79	y
111 1010	172	122	7A	z
111 1011	173	123	7B	{
111 1100	174	124	7C	|

Does this mean that a PDF document uses the same codes? Actually a PDF document contains neither decoration denoting fonts, boldface, italics nor plain text. Rather a PDF document contains 'Glyphs' or a collection of glyphs that display as the text you see. Commonly, each Glyph will also contain its own custom type of encoding for the letter 'C' or 'b' for example different from an ASCII or Unicode value.

In most cases, when standard encoding is used, each Glyph or collection will also contain the associated values or ASCII code needed to map and display the correct character during a conversion to Microsoft Word as shown in an example below.

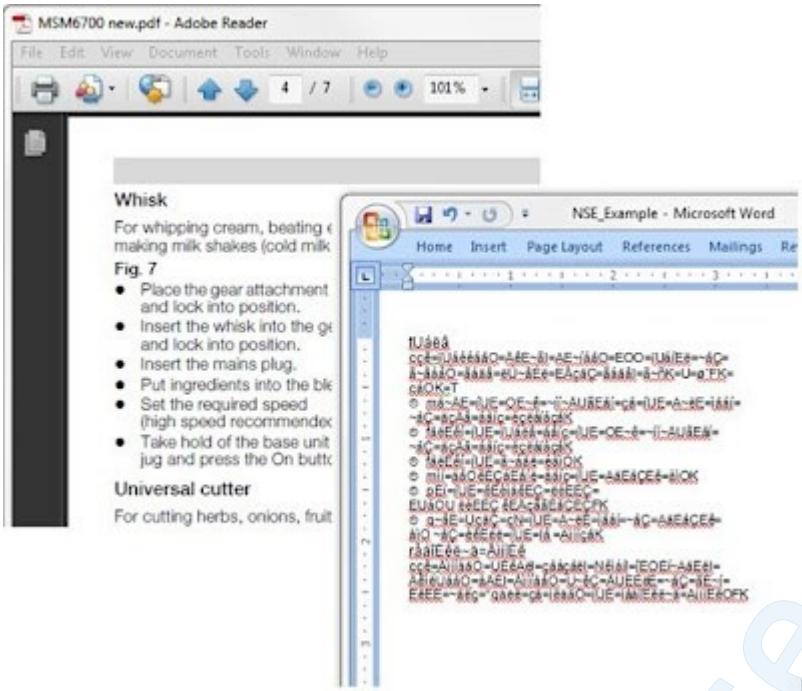
Glyph name;Unicode scalar value

Aring;00C5
Aringacute;01FA
Aringbelow;1E00
Aringsmall;F7E5
Asmall;F761
Atilde;00C3

Non-Standard Encoding

Unfortunately there are no rules for PDF creation utilities to require the use of standard encoding, like ASCII, standard glyph names or mapping of Glyph names to ASCII codes.

While the PDF may appear fine when viewed in Adobe Reader or Acrobat, the document actually lacks necessary encoding needed to convert successfully to Microsoft Word. For a better understanding please note the example below:



Notice the PDF looks great. How you can tell if a PDF has standard encoding or not, is to copy and paste the text from the PDF to Word - if Word can display the correct character then it is encoded, if not, it is non-standard encoding.

This is known as **Non Standard Encoding (NSE).**

While the created PDF document may render and look OK, the standard encoded associations to the Hexadecimal or Decimal values do not exist in the document making it impossible for Microsoft Word to map the values, associate and display the correct characters. You may have seen these results while working with a document created with low quality PDF creation tools or in a poorly scanned document.

So what happens when non standard encoding exists in a PDF document? It is then up to software such as Solid Converter PDF or Solid PDF Tools to determine and accurately recreate each character in Word. When NSE is detected the conversion engine of Solid Converter PDF and Solid PDF Tools work to study each character individually in a process of rebuilding the character or characters in order to supply Microsoft Word with the necessary encoding to display each character correctly. This process requires a high degree of recognition and reconstruction accuracy.

To continually study and improve our conversion engine, Solid Documents tests thousands of PDF files in a process using Solid Framework SDK to automate a rich set of conversion tests across a wide spectrum of PDF files in order to insure high quality and accurate conversion software.

With nearly decade of success in delivering best-in-class document reconstruction and archiving software, Solid Documents provides a technically sound and innovative document reconstruction software solution for standard as well as non-standard encoded PDF files.