Base32 Lowercase General Text
UTF-5 LG

(Sipendra Sinha, Vihar Innovations LLP)

INTRO

Each byte of human readble text consumes, 8 bits or 1 byte. Using 8 bits, enables a wide 256 character set (UTF-8). But case insensitive descriptional text (void of any numerals), need only 26 character set, space, and basic punctuation marks. All the requirements for such text can fit in 32 character set. This would require just 5 bits for representation of a single character. Thus using 37.5 % less space, than UTF-8 for such texts.

RATIONALE

The bitcoin's blockchain technology has enabled, permanent storage of arbitrary messages, in the computationally most secure data structure in the world. Blockchain's primary purpose is holding chain of financial transactions, so the storage of such data, has been reduced to only 40 bytes. Therefore, if a character set of 32 symbols is used, then, available 40 bytes could pack-in more data.

ALPHABETS

Given below is the representation of 32 symbols and their corresponding 5 bits.

BASE32 LG-TEXT

00000	space	01000	h	10000	p	11000	x
00001	a	01001	i	10001	q	11001	y
00010	b	01010	j	10010	r	11010	z
00011	c	01011	k	10011	s	11011	.
00100	d	01100	l	10100	t	11100	,
00101	e	01101	m	10101	u	11101	-
00110	f	01110	n	10110	v	11110	?
00111	g	01111	o	10111	w	11111	-

DATA CONVERSION TO BASE32 LG-TEXT

1. Divide the raw binary data into groups of 5 bits, and use the 32 symbols to convert to base32 lg-text.

2. If the number of bits is not a multiple of 5, then zero-pad the last remaining bits to conform to a 5 bit
group.

Example - If Binary Data = 0101 11011 = 01011 1011

Here the remaining 1011 must be padded as 01011
Making the Data as 01011 01011

Base32 LG-TEXT k k

DATA CONVERSION FROM BASE32 LG-TEXT

1. Concatenate 5 bit group with the next 5 bit group, take the first 8 bits, convert to utf-8 or hexadecimal represenation. Concatenate the remaining bits with the next group, then repeat the process.

Example - If base32 lg-text Data = a b c = 00001 00010 00011

For hex representation - 0000 1000 1000 011
0 a a

The remaining 011 is padded
as 0011 (hex - 2)

REFERENCES

[1] Human-oriented base-32 encoding
http://philzimmermann.com/docs/human-oriented-base-32-encoding.txt

[2] Base32 Encoding - Douglas Crockford
http://www.crockford.com/wrmg/base32.html

[3] UTF-5, a transformation format of Unicode and ISO 10646
https://tools.ietf.org/html/draft-jseng-utf5-01

[4] [RFC 4648] Base 32 Encoding
https://tools.ietf.org/html/rfc4648#page-8