Legal Codes

K. Ramanraj, M.L.
Advocate
High Court, Madras

26th February, 2010

Seminar on Science, Technology and Law
Department of Legal Studies
University of Madras

Daksinamurti Stotra

Obeisance to him Sri Daksinamurti, who is the Guru, who at the time of spiritual awakening, has verily realized his own Self, the one without a second, having understood that the world is within oneself even as a city reflected in a mirror is, but projected as if it is outside, by maya, as in dream. [1]

Obeisance to him, Sri Daksinamurti, who is the Guru,who, out of his free will, like the magician or a great yogi, manifests this world, which was, before creation, undifferentiated even as the sprout was within the seed, and became variegated later, on account of its association with space and time, brought forth by maya. [2]

Obeisance to him, Sri Daksinamurti, by whose brilliance, which is of the nature of existence, (this world which is) similar to unreality shines, who is enlightening those who have taken refuge in him by the message of the Vedas viz., 'Thou art verily that!, and by realizing whom, there is no return to this ocean of transmigration. [3]

Obeisance to him, Sri Daksinamurti, who is the Guru, whose consciousness is flowing out through the senses like the eyes etc., even as a powerful light kept within a pot full of holes (flowing though through the holes), following whom - the resplendent One - this whole world is shining and thinks, 'I know.' [4]

(Some) disputants who can be compared to the dull witted, being extremely deluded, think that the body, the vital airs, the senses, the fickle intellect and the void are the Atman. Obeisance to him, Sri Daksinamurti, who is the Guru, who dispels the great delusion that has been created by the play of the power of maya. [5]

Verses 6-10 ...

Obeisance to Sri Daksinamurti, who is the Guru, the Self, who in the deep sleep state induced by the withdrawal of the senses, being covered by maya - like the sun and the moon devoured by Rahu- was 'existence' only, and who at the time of waking, recognizes Himself as, '(It is I) who previously slept. [6]

Obeisance to Sri Daksinamurti, who is the Guru, who by the auspicious mudra is revealing to his votaries his own Self, which is persistently present as the 'I', always shining inside, in all the various and mutually exclusive states like childhood etc., as also waking etc. [7]

Obeisance to Sri Daksinamurti, who is the Guru, (who as) this person, being deluded by maya sees the world both in sleep and in the waking state, as (full of) differences (brought about by such) relationships as cause and effect, property and owner, disciple and teacher as also father, mother and so on. [8]

Obeisance to Sri Daksinamurti, who is the Guru, whose eightfold form is verily this world of the sentient and the insentient, comprising of earth, water, fire, air, sky, sun, moon and human being, and beyond whom - the gratest and the omnipresent - nothing else exists according to the discerning people. [9]

Since the principle of universal Self-hood has been revealed in this hymn, therefore, by listening to it and reflecting on its meaning as also by meditating on it and singing it, the attainment of identity with Isvara, together with the great power of being the universal Self, comes about automatically. Also, the unobstructed power that manifests itself in eight (different) ways is obtained. [10]

The highest purpose of codification

- Adi Sankara's ancient “Daksinamurthy Stotra” is remarkable in succintly summarising 'self-hood' that could be applied to machines as well
- Every single idea in the Stotra is capable of being implemented as program code
- A full set of legal codes used and applied by humans, in suitable form, would give life to AI systems
- This would also help to execute legal codes with speed, efficiency and accuracy

Codification:

In law, codification is the process of collecting and restating the law of a jurisdiction in certain areas, usually by subject, forming a legal code, i.e. a codex (book) of law.

-Wikipedia

Examples of codification

- Mitakshara was written by Vijnaneshwara during the reign of Vikramarka, a Chalukya ruler of 11th century A.D.
  - Its a commentary on the Yajnavalkya Smriti
  - Law, those days, consisted of treatises by jurists accepted as authoritative text due to tremendous scholarship, logical analysis and sheer force of intellect of the author.
  - Vijnaneswara abandoned the theory of connection through the rice-ball offering & accepted the theory of tranmission of constituent particles
  - The new definition was revolutionary – divested ”sapinda” of religious meaning
  - Cited from ”The importance of Mitakshara in the 21st Century” by Hon'ble Mr. Justice Markenday Katju, AIR 2005 Journal 215
- Indian Penal Code, Evidence Act, Criminal Procedure Code
- The Civil Procedure Code
- Codification by the Law Commission

Codification in the US

In the United States, acts of Congress, such as federal statutes, are published chronologically in the order in which they become law in official pamphlets called "slip laws," and are grouped together in official bound book form, also chronologically, as "session laws." The "session law" publication for Federal statutes is called the United States Statutes at Large. An act may be classified as either a "Public Law" or a "Private Law."

Because each Congressional act may contain laws on a variety of topics, many acts, or portions thereof are also rearranged and published in a topical, subject matter codification. The official codification of Federal statutes is called the United States Code. Generally, only "Public Laws" are codified. The United States Code is divided into "titles" (based on overall topics) numbered 1 through 50. Title 18, for example, contains many of the Federal criminal statutes. Title 26 is the Internal Revenue Code, the Bankruptcy Code in Title 11 of the United States Code, or the Judiciary Code in Title 28.

In the United States, the individual states, either officially or through private commercial publishers, generally follow the same three-part model for the publication of their own statutes: slip law, session law, and codification.

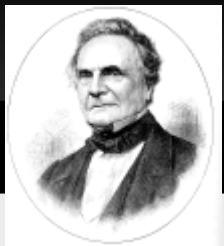
Code: In new senses

The noun 'code' has 3 senses in modern usage:

1. code, codification -- (a set of rules or principles or laws (especially written ones))
2. code -- (a coding system used for transmitting messages requiring brevity or secrecy)
3. code, computer code -- ((computer science) the symbolic arrangement of data or instructions in a computer program or the set of such instructions)

The verb code has 2 senses :

1. code -- (attach a code to; "Code the pieces with numbers so that you can identify them later")
2. encode, code, encipher, cipher, cypher, encrypt, inscribe, write in code -- (convert ordinary language into code; "We should encode the message for security reasons")



Babbage on An. Engine

In 1840 I received from my friend M. Plana a letter pressing me strongly to visit Turin at the then approaching meeting of Italian philosophers. In that letter M. Plana stated that he had inquired anxiously of many of my countrymen about the power and mechanism of the Analytical Engine. He remarked that from all the information he could collect the case seemed to stand thus:—

"Hitherto the legislative department of our analysis has been all powerful—the executive all feeble.

"Your engine seems to give us the same control over the executive which we have hitherto only possessed over the legislative department."

Considering the exceedingly limited information which could have reached my friend respecting the Analytical Engine, I was equally surprised and delighted at his exact prevision of its powers. Even at the present moment I could not express more clearly, and in fewer terms, its real object. I collected together such of my models, drawings, and notations as I conceived to be best adapted to give an insight into the principles and mode of operating of the Analytical Engine. On mentioning my intention to my excellent friend the late Professor MacCullagh, he resolved to give up a trip to the Tyrol, and join me at Turin.

Evolution of computer codes

- punch cards used by Jacquard, Babbage to code instructions to machines for execution
- Telegraph codes were popular and widely used
- Telegrams sent with pre-arranged codes resulted in valid and legally enforceable contracts
- ASCII (8 bits) evolved from Baudot 5 bit codes
- high level languages of today: c, c++, java ...
- scripting tools: php, perl, bash ...
- minimal turing complete code: forth [8k memory !]

Two English Cases:

D.P. Anderson & Co. Ltd. v. The Lieber Code Co.
[1917] 2 KB 469

A telegraphic code consisting of made words of five letters suitable for coding purposes, each of which was itself meaningless, and differed from every other word in at least two out of the five letters, was proper subject of copyright.

****

Ager v. Collinridge
(1886) 2 TLR 291

The defendant used many of the words listed in the standard telegram code but assigned their own meanings and numbers to the terms making them suitable to facilitate transmissions pertaining to timber trade. Copyright was found in the subject matter of Ager's ciphers and codes.

Bentley's Second Phrase

The "Bentley's Code Phrases"
[http://www.archive.org/details/bentleyscomplete00bentuoft]
first published in 1906 continued to be commonly used till the end of 1960's. That and other codes were widely used by commercial establishments.

"Coding" was popular then at grass root levels. If the 5 letter codes were well understood, popular and used widely, there are no reasons why the present coding done with computer languages based on 8 bit octets can't be used with the same ease by the general public.

Coding languages need to start circulating widely among the public the way Bentley's Code ruled from 1906 to the 1960s.

Bentley's Code: A sample

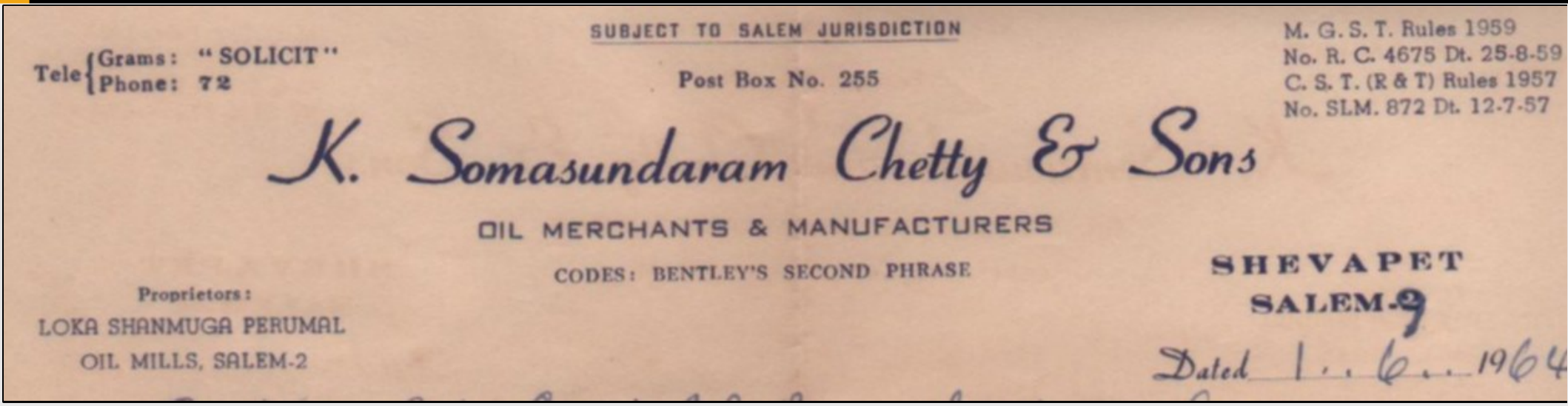
Genesis: In the late 19th and earlier 20th Centuries, there were Code Books created because telegram messages were charged by the word. As many as ten characters in a grouping were considered a word by the telegraph companies. Commercial Code Books, such as the Acme Code Words, or the Bentley's Complete Phrase Code were available to companies, enabling them to send complex messages in only a few "words."

Sample: For instance, if someone used Bentley's phrase book, he or she might choose the following letter groupings:

DIZUH (contracts for)
DAELF (computing)
FEAVO (equipment)
RUGUB (has/have been signed)
KUKIB (New York)
CUGYA (commence)
OKGAP (production)
ICSCO (immediately).

Thus, the message, DIZUHDAELF FEAVORIGUB KUKIBCUGYA OKGAPICSCO, four "words," would translate to, "Contracts for computing equipment have been signed [in] New York. Commence production immediately." This would be in place of 12 normal words (13 if the implied "in" is included); a savings of at least 75 percent. Of course, for someone without the Code Book, the message would be unreadable, but the message was sent primarily for economy, not security.

Code: The Past



SUBJECT TO SALEM JURISDICTION

Tele { Grams: "SOLICIT" / Phone: 72

Post Box No. 255

M. G. S. T. Rules 1959
No. R. C. 4675 Dt. 25-8-59
C. S. T. (R & T) Rules 1957
No. SLM. 872 Dt. 12-7-57

K. Somasundaram Chetty & Sons

OIL MERCHANTS & MANUFACTURERS

CODES: BENTLEY'S SECOND PHRASE

SHEVAPET
SALEM-9

Proprietors:
LOKA SHANMUGA PERUMAL
OIL MILLS, SALEM-2

Dated 1..6...1964

- Contract transmissions by means of telegraph from 1880's to 1960's used Codes - Std. Telegraph code to Bently's Second Phrase
- Many commercial letter heads acknowledged use of 'Bentley's Second Phrase' until late 1960's and faded from use after Telex and Fax
- Some of the old codes supported compression, error check and encryption but was manual and labour intensive
- Commercial 'codes' of the past are not very different modern coding languages in use of boilerplates and templates

Code: The Present & Future

- A Computer is a programmable machine that receives input, stores and manipulates data, and provides output in a useful format.
- The present day computing machines that receive code are capable of executing them, beyond merely printing out the codes.
- The Internet, since 1990's developed into a global system of interconnected computer networks that use the standard Internet Protocol Suite (TCP/IP) to serve billions of users worldwide.
- The Internet is a network of networks that consists of millions of private and public, academic, business, and government networks of local to global scope that are linked by a broad array of electronic and optical networking technologies.
- Computer programming (coding) is the process of writing, testing, debugging/troubleshooting, and maintaining the source code of computer programs. The process of writing source code often requires expertise in many different subjects, including knowledge of the application domain, specialized algorithms and formal logic.

Code: Machine & Human readable

- Legal Documents written in scripting languages like PHP
- Storage and access from public SQL servers
- Text standards: ASCII, Unicode
- Image and movie standards: PNG, MPEG ...
- Other standard specs: SQL, POSIX, HTML, ...

Government of India
Ministry of Communications & Information Technology
Department of Information Technology
Electronics Niketan
6, CGO Complex

New Delhi-110003
Dated: 27.11.2009

NOTIFICATION

Unicode 5.1.0

No. 2(32)/2009-EG-II: Whereas Department of Information Tecnology (DIT), Ministry of Communications and Information Technology, Government of India (GOI) is driving the National e-Governance Plan (NeGP) which seeks to create the right Governance and institutional mechanism and implement a number of Mission Mode Projects at the Centre and State Government; and

Whereas under NeGP, GOI is promoting the usage of Open Standards to avoid any technology lock-ins; and

Whereas Standards in e-Governance are a high priority activity, which will help ensure sharing of information and seamless interoperability of data across e-Governance applications. DIT, GOI has set up an Institutional Mechanism under NeGP to evolve/ adopt Standards for e-Governance; and

Whereas because of the lack of availability of information in local language there has been a slow progress in the Information and Communication Technology (ICT) sector and the benefits of ICT have not percolated down to the common man. Hence, "Localization and Language Technology" is an important area which is being addressed under Standardization; and

Whereas the Competent Authority on Standards has approved **Unicode 5.1.0** as Character Encoding Standard which is widely recognized all over the world for representation of multilingual text and also supports Indian languages as well as will ease Localization of applications for all the constitutionally recognized Indian languages; and

Therefore, Department of Information Technology, Government of India hereby notifies **Unicode 5.1.0** and its future versions as the Standard for e-Governance Applications w.e.f the date of notification.

(S.S. Rawat)
Joint Director

universal digital computer

0 represents false
1 represents true

Imagine an infinite tape on which 0's and 1's can be written or read

....10100110100001010101010111000...
^.....pointer to read, write, shift

universal digital computer

with a new feature: * legal rights management *

Do you have rights to read, write or shift?

Imagine an infinite tape on which 0's and 1's can be written or read

^.....pointer to read, write, shift

Permissions

```
-rw-r--r--      38450 2010-02-19 21:58 00_ML_IPR_Syllabus.pdf
-rw-r--r--     281924 2010-02-19 22:00 digital_signatures.pdf
-rw-r--r--    2598727 2010-02-19 21:59 freeculture.pdf
-rw-r--r--      67870 2010-02-19 21:57 privatisation.pdf
-rw-r--r--     649441 2010-02-19 22:10 tml_unicode.pdf
-rw-r--r--     473009 2010-02-19 21:57 go_can_read.pdf
-rwx------       1024 2010-02-17 18:00 go_cant_read.txt
```

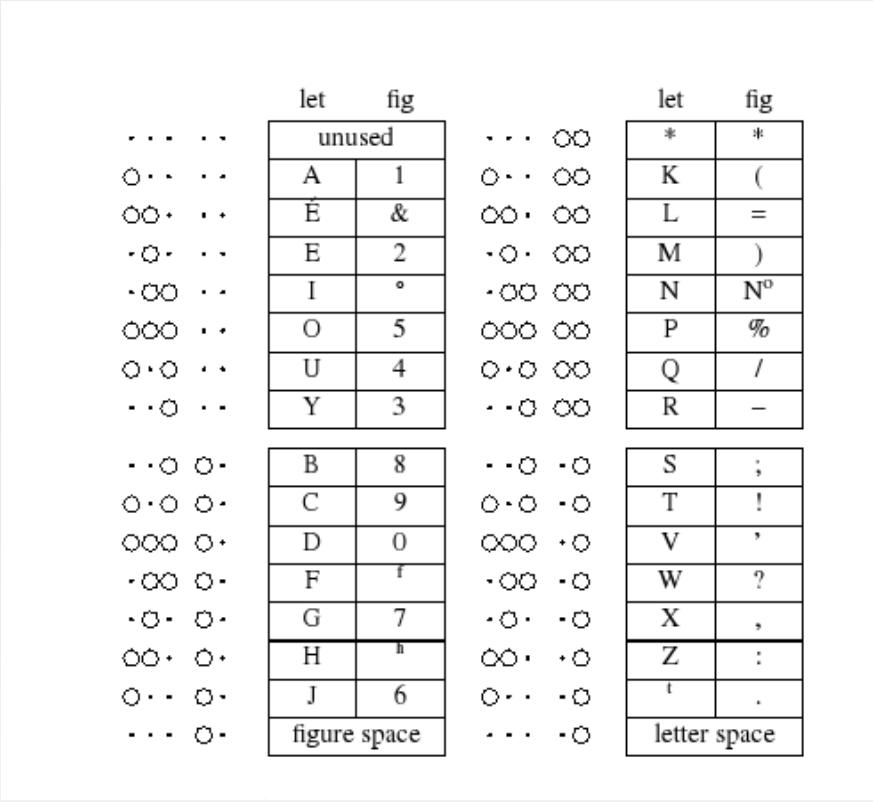
Persons::

U = You, the Creator/Author
G = Group
O = Others

Permissions::

r = read
w = write
x = execute/shift

Émile Baudot's five bit code



	let	fig
·· ·· ··	unused	
●· ·· ··	A	1
●● ·· ··	É	&
·● ·· ··	E	2
·●● ·· ··	I	°
●·· ·· ··	O	5
●·● ·· ··	U	4
··● ·· ··	Y	3
·· ·· ●·	B	8
●· ·· ●·	C	9
●●· ·· ●·	D	0
·●· ·· ●·	F	f
·●●· ·· ●·	G	7
●··· ·· ●·	H	h
●·●· ·· ●·	J	6
··· ·· ●·	figure space	

	let	fig
·· ·· ●●	*	*
●· ·· ●●	K	(
●● ·· ●●	L	=
·● ·· ●●	M	)
·●● ·· ●●	N	N°
●·· ·· ●●	P	%
●·● ·· ●●	Q	/
··● ·· ●●	R	–
·· ●· ●·	S	;
●· ●· ●·	T	!
●●· ●· ●·	V	’
·●· ●· ●·	W	?
·●●· ●· ●·	X	,
●··· ●· ●·	Z	:
●·●· ●· ●·	t	.
··· ●· ●·	letter space	

Five bit codes could accomodate 2^5, ie. 32 characters. Using “figure space” and “letter space” to shift, Baudot’s five bit code could represent 29 “figure” and “letter” characters as above

The ASCII table uses 2^8 characters. Chart showing 0-127

ASCII value	Character	Control character	ASCII value	Character	ASCII value	Character	ASCII value	Character
000	(null)	NUL	032	(space)	064	@	096	
001	☺	SOH	033	!	065	A	097	a
002	☻	STX	034	"	066	B	098	b
003	♥	ETX	035	#	067	C	099	c
004	♦	EOT	036	$	068	D	100	d
005	♣	ENQ	037	%	069	E	101	e
006	♠	ACK	038	&	070	F	102	f
007	(beep)	BEL	039	'	071	G	103	g
008	◘	BS	040	(	072	H	104	h
009	(tab)	HT	041	)	073	I	105	i
010	(line feed)	LF	042	*	074	J	106	j
011	(home)	VT	043	+	075	K	107	k
012	(form feed)	FF	044	,	076	L	108	l
013	(carriage return)	CR	045	-	077	M	109	m
014	♫	SO	046	.	078	N	110	n
015	☼	SI	047	/	079	O	111	o
016	►	DLE	048	0	080	P	112	p
017	◄	DC1	049	1	081	Q	113	q
018	↕	DC2	050	2	082	R	114	r
019	!!	DC3	051	3	083	S	115	s
020	π	DC4	052	4	084	T	116	t
021	§	NAK	053	5	085	U	117	u
022	▬	SYN	054	6	086	V	118	v
023	↨	ETB	055	7	087	W	119	w
024	↑	CAN	056	8	088	X	120	x
025	↓	EM	057	9	089	Y	121	y
026	→	SUB	058	:	090	Z	122	z
027	←	ESC	059	;	091	[	123	{
028	(cursor right)	FS	060	<	092	\	124	|
029	(cursor left)	GS	061	=	093	]	125	}
030	(cursor up)	RS	062	>	094	^	126	~
031	(cursor down)	US	063	?	095	_	127	⌂

Representing everything with ON & OFF

- ON-OFF states represent boolean values
  - `OFF represents 0 or FALSE`
  - `ON  represents 1 or TRUE`
- Each 0 or 1 is a "binary digit" or "bit" of information
- A BYTE (BinarY TablE) is a contiguous sequence of a fixed number of bits which has come to mean 8 bits "octet" capable of holding 256 values from 00000000 to 11111111
- ASCII – American Standard Code for Information Interchange character encoding based on the English Alphabet is the widely used standard
- The 95 printable ASCII characters are:
  !"#$%&'()*+,-./0123456789:;<=>?
  @ABCDEFGHIJKLMNOPQRSTUVWXYZ
  [\]^_`abcdefghijklmnopqrstuvwxyz{|}~
- `Source code by programmers is converted to machine code`

ASCII Chart

binary	glyph
0011 0000	0
0011 0001	1
0011 0010	2
0011 0011	3
0011 0100	4
0011 0101	5
0011 0110	6
0011 0111	7
0011 1000	8
0011 1001	9
0100 0001	A to
0101 1010	Z ..

Boolean Logic

AND

∧	0	1
0	0	0
1	0	1

OR

∨	0	1
0	0	1
1	1	1

NOT

a	0	1
¬a	1	0

ASCII Chart :::::::::::::::::::::::::::: Intelligence

ASCII Code **Most Significant Bits [MSB]**

LSB	**000**	**001**	**010**	**011**	**100**	**101**	**110**	**111**
0000	NUL,^@	DLE,^P	spc	0	@	P	`	p
0001	SOH,^A	DC1,^Q	!	1	A	Q	a	q
0010	STX,^B	DC2,^R	“	2	B	R	b	r
0011	ETX,^C	DC3,^S	#	3	C	S	c	s
0100	EOT,^D	DC4,^T	$	4	D	T	d	t
0101	ENQ,^E	NAK,^U	%	5	E	U	e	u
0110	ACK,^F	SYN,^V	&	6	F	V	f	v
0111	BEL,^G	ETB,^W	'	7	G	W	g	w
1000	BS,^H	CAN,^X	(	8	H	X	h	x
1001	HT,^I	EM,^Y	)	9	I	Y	i	y
1010	LF,^J	SUB,^Z	*	:	J	Z	j	z
1011	VT,^K	ESC,^[	+	;	K	[	k	{
1100	FF,^L	FS,^\	,	<	L	\	l	|
1101	CR,^M	GS,^[	-	=	M	]	m	}
1110	SO,^N	RS,^^	.	>	N	^	n	~
1111	SI,^O	US,^_	/	?	O	-	o	DEL

ASCII integers are converted to binary integers by flipping bits 5 & 4 to 0

Uppercase alphabetical characters are converted to lowercase by flipping bit 5 from 0 to 1

Uppercase characters are converted to the equivalent control characters by flipping bit 6 (msb) from 1 to 0

Church-Turing Thesis

- According to the Church–Turing thesis, a computer with a certain minimum threshold capability is in principle capable of performing the tasks of any other computer.

- A Turing machine has only a single data structure, a variable-length linear array called the tape. Each component of the tape contains just a single character.

- **....1000110100101100110110100101111000....**
  .^. --->read/write/shift pointer

- Any computable function can be computed by a Turing machine

- It takes almost no machinery to achieve universality, other than some sort of unlimited storage capacity. Even an extremely simple set of data structures and operations are sufficient to allow any computable function to be expressed.

- Anything can be done in LISP, Python, PHP, C...
  The differences between programming languages is not quantitative but qualitative in how elegantly, easily, and effectively things can be done

- Computers with capabilities ranging from those of a personal digital assistant to a supercomputer may all perform the same tasks, as long as time and memory capacity are not considerations.

- The same computer designs may be adapted for tasks ranging from processing company payrolls to controlling unmanned spaceflights.

Law at the core of computing: teaching machines to respect human rights

- Humans and other life forms are endowed with natural computing abilities.
- If we admit the Church-Turing thesis, in theory, all our computing functions could be performed by a computer.
- But then, why is it that common sense reasoning is not yet possible and the AI problem is without a solution?
- How would a robot know how to deal with humans and others?
- Assimov's three laws of robotics or are too simplistic - Law is more detailed in describing such matters and the best judge of what is relevant and what is not.
- The computing field has not taken law seriously enough, and that has prevented the evolution of robust AI systems.
- Porting the rules relating to the legal system, language, computing, arithmetic, vision, and other fields of knowledge would give computers a chance to do common sense reasoning.

Program structure theorem

- Bohm & Jacobini [1966] showed that any prime program could be written using only:
  - while statements; and
  - if statements
- This means, legal codes could be written with a minimal set of statement structures

Current coding efforts:

- www.mca.gov.in
- income-tax application processing
- eFilings before authorities
- eGovernance
- calpp-10.7
  - computer aided legal procedures and proceedings
  - .^.i goals
  - Fair Rent Calculator (Tamil Nadu)

Coding Standards

- Vendor neutral
- Free, Open, non-proprietary
  - “Open Document Format” by Justice Yatindra Singh at http://kvtrust.blogspot.com/2007/04/open-document-format.html
- Appropriate variable names
- Permissions
- Privacy
- Tools

Standards and Tools

The Free Software licenses give freedom
- to run the program, for any purpose
- to modify the program to suit your purpose
- to redistribute copies, either gratis or for a fee
- to distribute modified versions of the program
- to access source code to effectively exercise the above rights

A wide variety of Operating Systems under different licenses:
- GNU/Linux, GNU/HURD released under the GPL
- FreeBSD, NetBSD, OpenBSD *BSD under BSD license
- Plan 9, Open Solaris etc under their respective free licenses

Applications to suit many requirements:
- Shells: bash, sh, tch etc
- Programming Languages: gcc, perl, python, php, javascript, ..
- Web Servers: Zope, Apache ...
- Database Servers: PostgreSQL, MySQL ...
- Office Suites: Koffice, OpenOffice, ...
- Text editors, ide's : emacs, vi, eclipse
- Browsers and Mail : Firefox, Mozilla,...
- Security: OpenSSH
- Graphic tools: Gimp, Inkscape, ImageMagic,
- Others: BIND, Sendmail, postfix, GNU Mailman ...

Free Software &
Open Standards

{Vendor Neutral}

IEEE - POSIX
ANSI C, SQL
W3C
- HTML
- XML
- DOM
- CSS

NSA HTTP 1.1

Legal codes: Roles

- Governments/Authorities:
  - Transparency and openness
  - Debates and discussions with public
- Experts/Public
  - Critical appraisal of projects
  - Code contributions and suggestions
- Universities/Research Institutes
  - Study impact of legal codes
  - Research future projects
  - Teach lessons from the past

A lesson from porting Indic scripts to Unicode

- While most indic scripts require 200+ characters, request to Unicode to allot just 127 characters for each indic script was a costly error.
- While China requested 40,000+ code points for their ideographs, India failed to request the requisite code points for our scripts
- The Unicode planes 1 to 3 have been fully utilized with other languages allotted code points, and it is hard to change now
- As a result, upto 9 bytes may be required to represent a single character like “கை” in Tamil against a mere 3 bytes if allocation had been made for the full complement of characters
- A detailed look:

Unicode – the emerging encoding standard

Unicode is an industry standard designed to allow text and symbols from all of the writing systems of the world to be consistently represented and manipulated by computers. [Wikipedia]

- Unicode consists of:
  - Character repertoire
  - Encoding methodology
  - Set of standard character encodings
  - Sets of code charts for visual reference,
  - Enumeration of character properties
  - Rules for normalization, decomposition, collation and rendering.

- Unicode features:
  - Range 0000 to 10FFFF
  - 65535 chars in the Basic Multilingual Plane 0 range from U+0000 to U+FFFF
  - Supplementary Planes 1 to 16 for code points U+10000 to U+10FFFF
  - Script based encoding
  - Each character is assigned a unique code point that is an integer value between 0 and 1114112

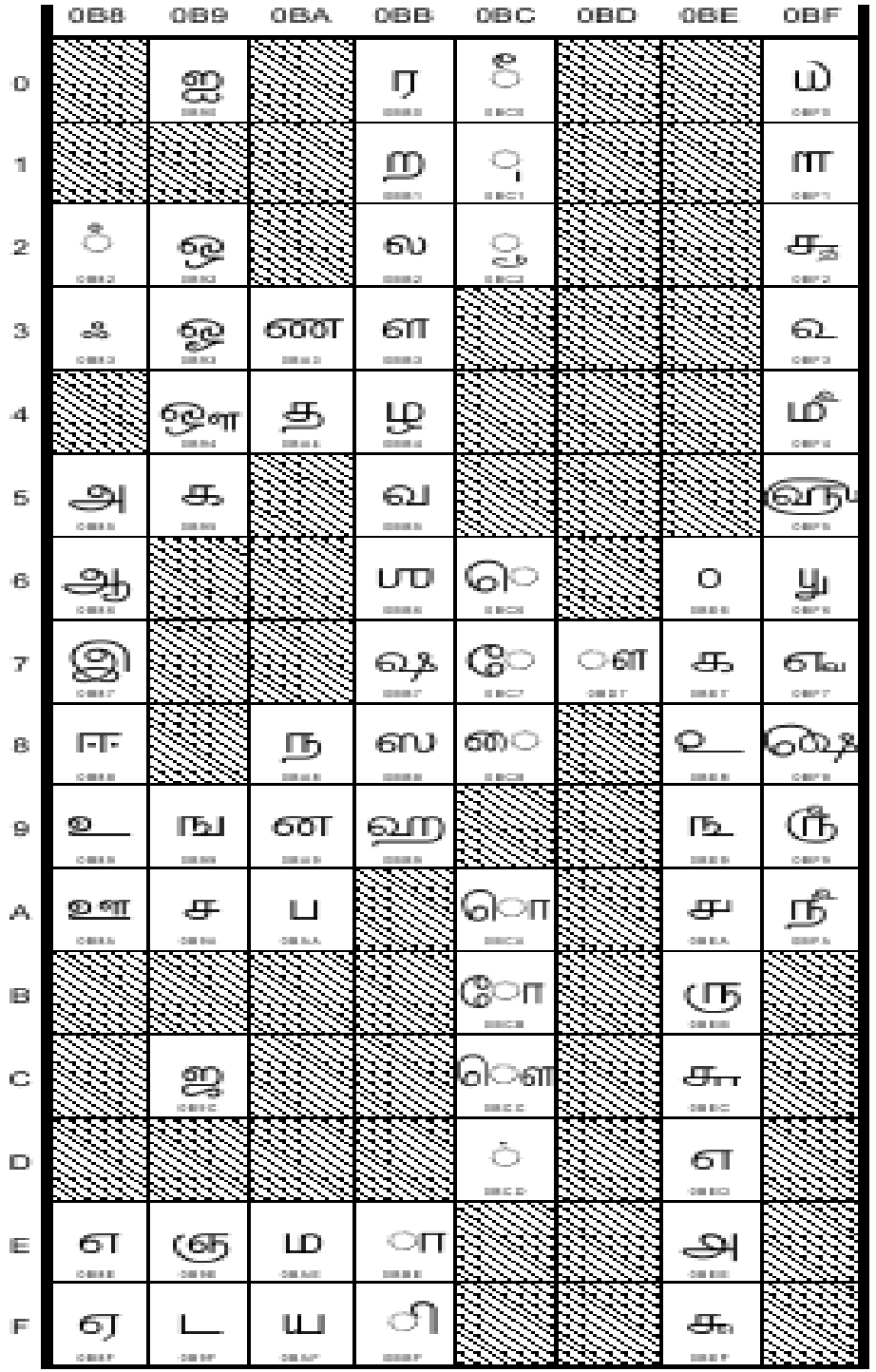
Unicode: Supported Indic Scripts

- # Property: Block
  - 0000..007F; Basic Latin
  - 0900..097F; Devanagari
  - 0980..09FF; Bengali
  - 0A00..0A7F; Gurmukhi
  - 0A80..0AFF; Gujarati
  - 0B00..0B7F; Oriya
  - 0B80..0BFF; Tamil
  - 0C00..0C7F; Telugu
  - 0C80..0CFF; Kannada
  - 0D00..0D7F; Malayalam

All languages in the VIII Schedule of the Constitution of India are supported, except Manipuri (item 9)

தமிழ் யுனிக்கோட்

- B80-BFF Range
  - (128 :: 2944-3071)
- Utilisation:
  - மொத்தம்71
    - உயிர் 12
    - மெய் 18
    - ஆய்தம் 1
    - இதர குறிகள் 40
  - காலி இடம் 57



Unicode Collation Algorithm :: DUCET

UCA Tamil Collation Chart

source:

- http://developer.mimer.com/charts/UCA_tamil.htm
- http://www.unicode.org/Public/UCA/4.1.0/allkeys.txt

Issues with Unicode Tamil Range B80-BFF

- Sort order not given correctly in DUCET
- Expensive encoding for Tamil chars
  - Three to twelve bytes to store chars
  - Databases could take three times more space
  - Heavy burden on a light language
- Missing chars
  - No code points for உயிர் மெய்
  - Fraction and measurement symbols missing
- Rigid Scheme
  - Brahmi -> Vatteluthu -> ??? Evolution ???

தமிழ் எழுத்துக்கள் - TUNE

(Font : TAUN_Elango_Barathi)

UNICODE_NEW

	E20	E21	E22	E23	E24	E25	E26	E27	E28	E29	E2A	E2B	E2C	E2D	E2E	E2F	E30	E31	E32	E33	E34	E35	E36	E37	E38	E39	E3A	E3B	E3C	E3D	E3E	E3F
0	Null	க்	ங்	ச்	ஞ்	ட்	ண்	த்	ந்	ப்	ம்	ய்	ர்	ல்	வ்	ழ்	ள்	ற்	ன்	ஜ்	ஶ்	ஷ்	ஸ்	ஹ்	க்ஷ்					ணு	உத	
1	அ	க	ங	ச	ஞ	ட	ண	த	ந	ப	ம	ய	ர	ல	வ	ழ	ள	ற	ன	ஜ	ஶ	ஷ	ஸ	ஹ	க்ஷ					று	சு	
2	ஆ	கா	ஙா	சா	ஞா	டா	ணா	தா	நா	பா	மா	யா	ரா	லா	வா	ழா	ளா	றா	னா	ஜா	ஶா	ஷா	ஸா	ஹா	க்ஷா					ணூ	பீ	
3	இ	கி	ஙி	சி	ஞி	டி	ணி	தி	நி	பி	மி	யி	ரி	லி	வி	ழி	ளி	றி	னி	ஜி	ஶி	ஷி	ஸி	ஹி	க்ஷி					உ	சூ	
4	ஈ	கீ	ஙீ	சீ	ஞீ	டீ	ணீ	தீ	நீ	பீ	மீ	யீ	ரீ	லீ	வீ	ழீ	ளீ	றீ	னீ	ஜீ	ஶீ	ஷீ	ஸீ	ஹீ	க்ஷீ					ணை	சு	
5	உ	கு	ங	சு	ஞு	டு	ணு	து	நு	பு	மு	யு	ரு	லு	வு	ழு	ளு	று	னு	ஜு	ஶு	ஷு	ஸு	ஹு	க்ஷு					லை	ப	
6	ஊ	கூ	ஙூ	சூ	ஞூ	டூ	ணூ	தூ	நூ	பூ	மூ	யூ	ரூ	லூ	வூ	ழூ	ளூ	றூ	னூ	ஜூ	ஶூ	ஷூ	ஸூ	ஹூ	க்ஷூ					ளை	பூ	
7	எ	கெ	ஙெ	செ	ஞெ	டெ	ணெ	தெ	நெ	பெ	மெ	யெ	ரெ	லெ	வெ	ழெ	ளெ	றெ	னெ	ஜெ	ஶெ	ஷெ	ஸெ	ஹெ	க்ஷெ					னை	நி	
8	ஏ	கே	ஙே	சே	ஞே	டே	ணே	தே	நே	பே	மே	யே	ரே	லே	வே	ழே	ளே	றே	னே	ஜே	ஶே	ஷே	ஸே	ஹே	க்ஷே					○	நூ	
9	ஐ	கை	ஙை	சை	ஞை	டை	ணை	தை	நை	பை	மை	யை	ரை	லை	வை	ழை	ளை	றை	னை	ஜை	ஶை	ஷை	ஸை	ஹை	க்ஷை					●	வ	
A	ஒ	கொ	ஙொ	சொ	ஞொ	டொ	ணொ	தொ	நா	பொ	மொ	யொ	ரொ	லொ	வொ	ழொ	ளொ	றொ	னொ	ஜொ	ஶொ	ஷொ	ஸொ	ஹொ	க்ஷொ					★	ரி	
B	ஓ	கோ	ஙோ	சோ	ஞோ	டோ	ணோ	தோ	நோ	போ	மோ	யோ	ரோ	லோ	வோ	ழோ	ளோ	றோ	னோ	ஜோ	ஶோ	ஷோ	ஸோ	ஹோ	க்ஷோ					ஈ	உ	
C	ஔ	கௌ	ஙௌ	சௌ	ஞௌ	டௌ	ணௌ	தௌ	நௌ	பௌ	மௌ	யௌ	ரௌ	லௌ	வௌ	ழௌ	ளௌ	றௌ	னௌ	ஜௌ	ஶௌ	ஷௌ	ஸௌ	ஹௌ	க்ஷௌ					ஓ	சூ	
D	ஃ																								ஸ்ரீ					றை		
E																														வ		
F																														?		

Note : E39, E3A, E3B & E3C are reserved for TAMIL accent, Punctuation marks & any feature characters.
E3D to E3F are reserved for Tamil Symbols

TUNE

- The GOOD
  - Code points for almost all characters
  - Symbols for Fractions included
  - Table in natural sort order (almost)
- The BAD
  - Encoding in Unicode Private Use Area Range E000–F8FF
    - Reason enough to reject TUNE
  - DUCET Sort order ???
  - Archaic orthographical Tamil chars missing
  - Measurement symbols missing

Unicode Private Use Area Range: E000–F8FF

- The Private Use Area does not contain any character assignments, consequently no character code charts or namelists are provided for this area.
- Q. What about using private-use characters for encoding Tamil clusters?
  - A: This is a fine solution for internal processing, if an alternate representation is useful for the particular process. For example, a text-to-speech program might use a private-use encoding for English, whereby letters were separated according to pronunciations -- the 'o' in 'love', 'rove', and 'move' all getting different private-use characters.
  - However, such implementations have limited usage. Private-use characters may overlap between different implementations, so general purpose programs cannot assume any particular interpretation of such characters. In general interchange, such as in search engines, private-use characters are typically treated as unknown characters or ignored. As a result private-use characters are inappropriate for open interchange.
- Source: http://www.unicode.org Tamil FAQ and PUA pages

Unicode Planes

* Plane 0 (0000–FFFF): Basic Multilingual Plane (BMP)
* Plane 1 (10000–1FFFF): Supplementary Multilingual Plane (SMP)
* Plane 2 (20000–2FFFF): Supplementary Ideographic Plane (SIP)
* Planes 3 to 13 (30000–DFFFF) are unassigned
* Plane 14 (E0000–EFFFF): Supplementary Special-purpose Plane (SSP)
* Plane 15 (F0000–FFFFF) reserved for the Private Use Area (PUA)
* Plane 16 (100000–10FFFF), reserved for the Private Use Area (PUA)

Notes:

- 40k Chinese ideographs in SIP (Plane 2)
- Chinese language uses a logographic script — that is, a script where one or two "characters" corresponds roughly to one "word" or meaning
- GB18030 is the Chinese National Standard for information interchange, the Chinese equivalent of UTF-8 and as of August 1, 2006, support for this character set is officially mandatory for all software products sold in the PRC.
- Indic scripts and root words may seek Plane 3 and enforce it.

The Encoding Challenge

- ¼ = கால் 
- ½ = அரை
- 1/8 = அரைக்கால்
- ¾ = முக்கால் 
- 1/16 = வீசம்
- 1/80 = காணி
- 1/320 = முந்திரி
- 1/20 = மா
- 1/5 = ?
  - நாலுமா
- 3/20 = ?
  - மும்மா
- 1/10 = ?
  - இருமா
- 1/40 = ?
  - அரைமா
- 1/640 = கீழ்ழரை
- 1/5120 = கீழ்வீசம்
- 1/2560 = கீழ்ழரைக்கால்
- 1/102400 = கீழ்முந்திரி
- 1/1075200 = இம்மி
- 8 = எட்டு = byte = ?
- 16 = ?
- 256, 512, 1024 .... ??

How to encode all this information ?

யுனிக்கோட் எண்வகை

நன்னூல் நூற்பா 193 : எட்டு என்பதற்கு சிறப்பு விதி

"எட்டன் உடம்புணவ் வாகும் என்ப"

விளக்கம்: இறுதி உயிர்மெய் கெட்டு நின்ற எட்டு என்னும் எண்ணினது டகர மெய் நாற்கணமும் வரின் ணகர மெய்யாகத் திரியும்

எட்டு + ஆயிரம் = எண்ணாயிரம்

எட்டு + வகை = எண்வகை

எட்டு + நாள் = எண்ணாள்

எட்டு + கலம் = எண்கலம்

UTF-8 = யுனிக்கோட் எண்வகை மாற்று

UTF-16 = யுனிக்கோட் ஈரெண்வகை மாற்று

UTF-32 = யுனிக்கோட் நாலெண்வகை மாற்று

UTF-8 யுனிக்கோட் எண்வகை மாற்று

Unicode Transformation Format :: UTF-8 :: Code range

hexadecimal	**UTF-8**
000000 – 00007F 128 codes	0zzzzzzz
000080 – 0007FF 1920 codes	110yyyyy 10zzzzzz
000800 – 00FFFF 63488 codes	1110xxxx 10yyyyyy 10zzzzzz
010000 – 10FFFF 1048576 codes	11110www 10xxxxxx 10yyyyyy 10zzzzzz

- Notes:
  - The first 128 characters (Basic Latin) need only one byte for encoding.
  - The next 1920 characters need two bytes to encode.
  - This includes Latin alphabet characters with diacritics, Greek, Cyrillic, Coptic, Armenian, Hebrew, and Arabic characters.
  - The rest of the BMP characters use three bytes, and additional characters are encoded in four bytes. [Wikipedia on UTF-8]

References

- http://www.unicode.org/reports/tr10/
- http://unicode.org/Public/cldr/1.4.0/
- http://unicode.org/faq/tamil.html
- http://www.unicode.org/reports/tr6/
- http://people.w3.org/rishida/scripts/pickers/tamil/
- http://unifont.org/fontguide/
- http://www.tunerfc.tn.nic.in/
- http://developer.mimer.com/charts/tamil.htm
- நன்னூல் எழுத்ததிகாரம், நன்னூல் சொல்லதிகாரம்
- தொல்காப்பியம்
- கொடுந்தமிழ் - Joseph Beschi

Legal Coding: Sincere efforts needed

- The Unicode lessons:
  - There ought to be wider participation in the evolution of standards
  - Emerging trends in hardware and software should be more vigorously studied, researched and implemented as soon as possible
  - India has an ancient scientific culture that nurtured philosophy – that should be encouraged
  - Merit based order in matters pertaining to coding standards should be promoted

Legal Codes: Plan for the future

- At the moment, there is hardly any body of machine executable legal code worth talking
- The legal community should acquire skills in programming and attempt to port law to computers through substantial research
- If the legal community fails here, it may have disastrous consquences for all
- The use of machine executable legal codes at grass root levels should be encouraged and promoted

References

- Vincent P. Heuring & Harry F. Jordan, “Computer Systems Design and Architecture”, Prentice Hall, 2004
- Terrence W. Pratt & Marvin V. Zelkowitz, “Programming Languages”, Prentice Hall, 2001
- Eric Fischer, “The Evolution of Character Codes, 1874-1968”
- www.wikipedia.org; www.google.com; www.unicode.org
- Justice Markenday Katju, “The importance of Mitakshara in the 21st Century”, AIR 2005 Journal 215
- Justice Yatindra Singh, “Open Document Format” published online at http://kvtrust.blogspot.com/2007/04/open-document-format.html
- Charles Babbage, “Passages from the Life of a Philosopher”, 1840, published online at http://www.fourmilab.ch/babbage/lpae.html
- Swami Harshananda, “Daksinamurti Stotra with Manasollasa”, Ramakrishna Math, 1992
- K. Ramanraj, “Tamil at Unicode” presentation at ILUGC, on 9th December 2006