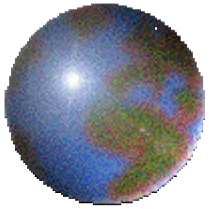
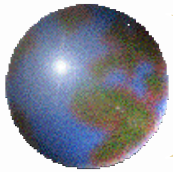


UNICODE™

Dr. Bruce K. Haddon

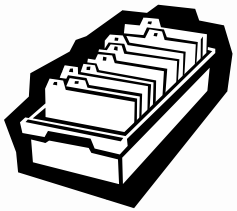
Paladin Software International

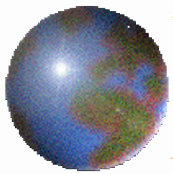
INCORPORATED



Character Encodings

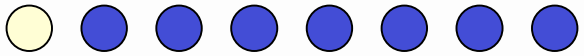
- Morse Code ●●●▬ ●●● ●▬●
- Baudot Code
- Hollerith
- ASCII
- EBCDIC
- *etc.*



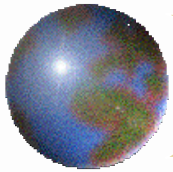


ASCII (ANSI-X3.4)

- Standard defined **1963, 1968, 1986, 1997**
- 7-bit code

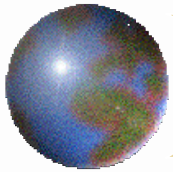


- Purpose: information interchange
- Popular choice for programming languages (*e.g.*, C, Ada, Java, *etc.*)
- Became the *de facto* code set and encoding for (too?) many applications.



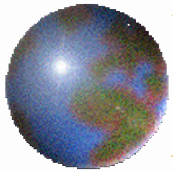
ASCII—The Code

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



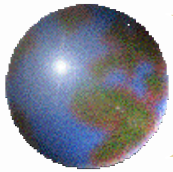
ISO 646

- International Standards Organization version of "ASCII"
- 7-bit codes
- Currently 25 National variants
  - (changes certain characters, *e.g.*, $5B_{16}$ "[" in ASCII is "Æ" in 646-DK)

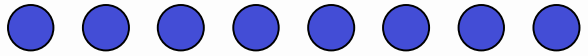


ISO 646—The Code

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

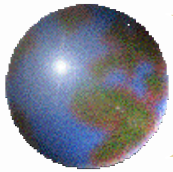


ISO/IEC 8859

- 8 bit codes 
- Currently, 16 variants (called “Parts”)
- 7-bit subset of each ≡ ASCII (exactly)
- Each 8859 variant redefines the code points from 80_{16}-FF_{16}
- *e.g.*, IS0/IEC 8859-1 is “Latin-1”,
  IS0/IEC 8859-5 is “Latin/Cyrillic”
  IS0/IEC 8859-9 is “Latin-5”
  IS0/IEC 8859-15 is “Latin-9” or “*Latin-0*”

ISO/IEC 8859-1—The "Latin-1" Code

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
8	XXX	XXX	BHP	NBH	IND	NEL	SSA	ESA	HTS	HTJ	VTS	PLD	PLU	RI	SS2	SS3
9	DCS	PU1	PU2	STS	CCH	MW	SPA	EPA	SOS	XXX	SCI	CSI	ST	OSC	PM	APC
A	NBSP	¡	¢	£	€	¥	|	§	¨	©	ª	«	¬	SHY	®	
B	°	±	?	?	´	µ	¶	·	¸	?	º	»	?	?	?	¿
C	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
D	D	Ñ	Ò	Ó	Ô	Õ	Ö	?	Ø	Ù	Ú	Û	Ü	Y	?	ß
E	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
F	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	?	?	ÿ

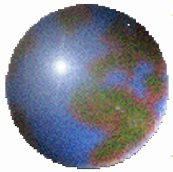


ISO/IEC 8859-15—The "Latin-9" Code

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
8	XXX	XXX	BHP	NBH	IND	NEL	SSA	ESA	HTS	HTJ	VTS	PLD	PLU	RI	SS2	SS3
9	DCS	PU1	PU2	STS	CCH	MW	SPA	EPA	SOS	XXX	SCI	CSI	ST	OSC	PM	APC
A	NBSP	¡	¢	£	€	¥	Š	§	š	©	ª	«	¬	SHY	®	
B	°	±	?	?		µ	¶	·		?	º	»	Œ	œ	Ÿ	¿
C	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
D	D	Ñ	Ò	Ó	Ô	Õ	Ö	?	Ø	Ù	Ú	Û	Ü	Y	?	ß
E	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
F	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	?	?	ÿ

Other "National" and ISO Standards

- Japan Industry Standards
  - Series of encodings (>15), all including ASCII and "wide character" ASCII
- Big 5 (Chinese)
- Korean
- UNIX Extended Code (UEC)
  - Escaping convention to allow intermixing of ASCII and any of the above (Open Consortium, OSF, UI, USLP: 1991)
- ISO-2022-JP, -JP1, -JP2, -CN, -CN EXT, -KP, -KR, -VN

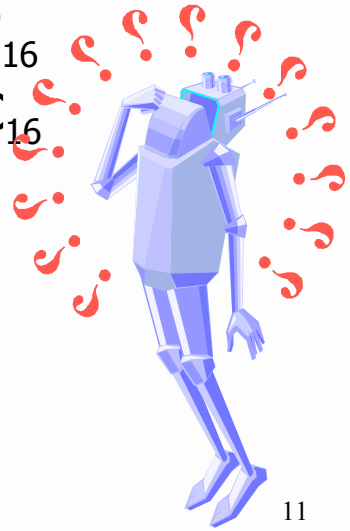


Shift-JIS

- ASCII
  - $21\text{-}7E_{16}$
- half-width katakana
  - $A1\text{-}DF_{16}$
- JIS X 0208:1977
  - 1st byte $81\text{-}9F_{16}$, $E0\text{-}EF_{16}$
  - 2nd byte $40\text{-}7E_{16}$, $80\text{-}FC_{16}$

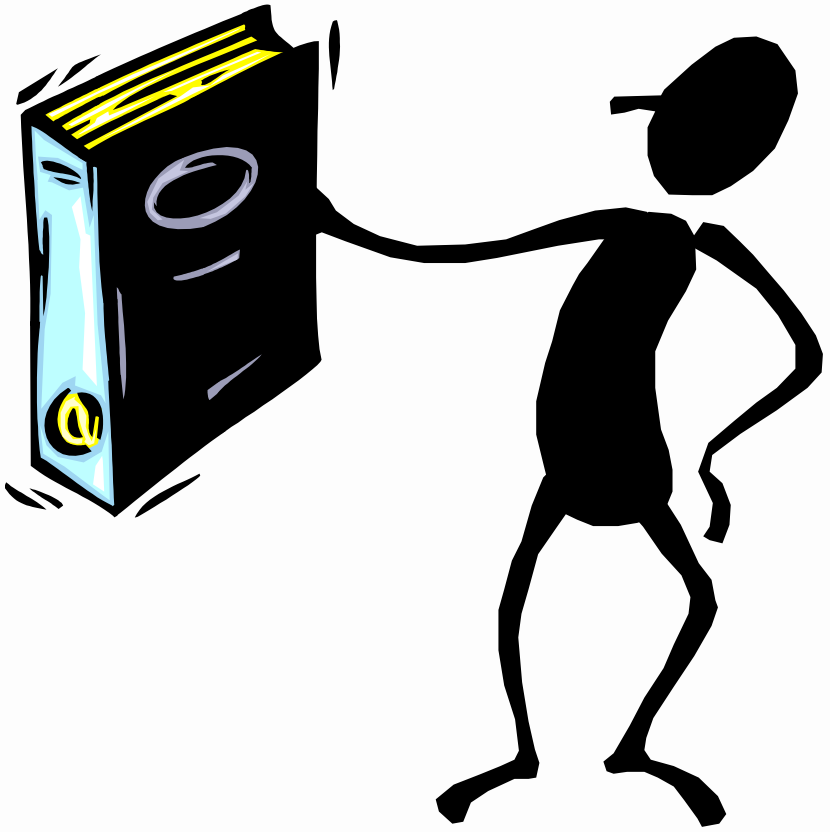
EUC-JP

- ASCII or JIS-Roman
  - $21\text{-}7E_{16}$
- half-width katakana
  - $8E_{16}$ followed by $A1\text{-}DF_{16}$
- JIS X 0208:1977
  - 1st byte $81\text{-}9F_{16}$, $E0\text{-}EF_{16}$
  - 2nd byte $40\text{-}7E_{16}$, $80\text{-}FC_{16}$
- JIS X 0212:1990
  - $8F_{16}$ followed by:
  - 2nd byte $A1\text{-}FE_{16}$
  - 3rd byte $A1\text{-}FE_{16}$



Terminology (1)

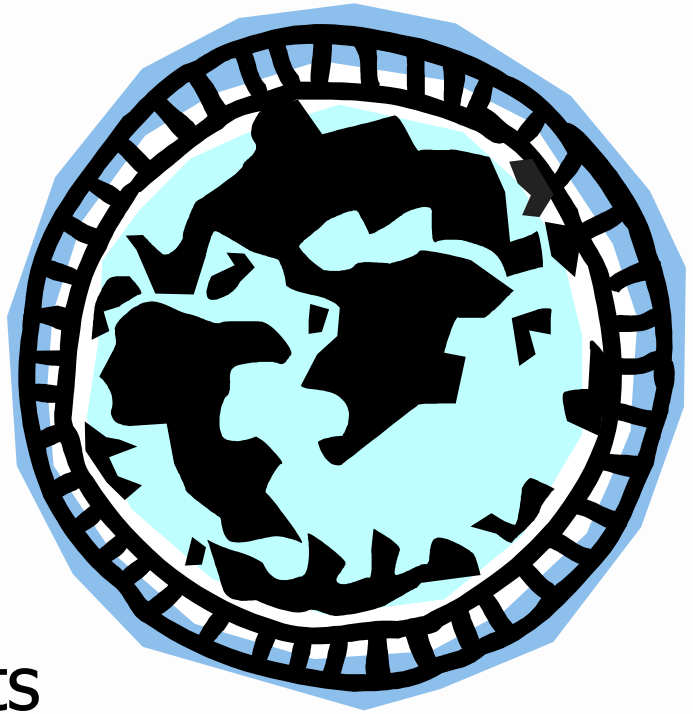
- “Character Set”
- “Glyph”
- “(Natural) Encoding”
  - “Code page/set”
- “Code point”
- “Transcoding”
- “Transformation”



Terminology (2)

- “Single byte, simple”
- “Double byte (simple)”
  - “Multi-byte (simple)”
- “Single byte, complex”
- “Bi-Directional” (‘bi-di’)
- “Universal”

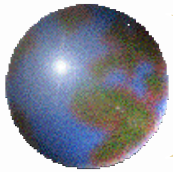
ISO/IEC-10646



- “Universal” character set
- Each code point is 32 (31 bits—first 0) (UCS-4)
- Initial approach, use “planes,” each containing defined national subsets
- 15 bits define “plane”, 16 bits define character encoding within plane.

The Unicode Standard

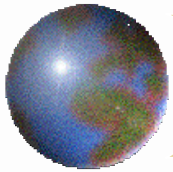
- Consortium of, now, 24 “full” members and 28“associate” members
- Interoperability with ISO 8859-1 Latin-1 (including ASCII)
- Encompassing all scripts in use—***now,*** all scripts ever used (or shall be used!)



The Unicode Consortium®

- Adobe Systems, Inc.
- Apple Computer, Inc.
- Basis Technology Corporation
- Compaq Computer Corporation
- Hewlett-Packard Company
- Hyperion Solutions
- IBM Corporation
- India, Ministry of Information Technology
- Justsystem Corporation
- Microsoft Corporation
- NCR Corporation
- Oracle Corporation
- Pakistan, National Language Authority
- PeopleSoft, Inc.
- Progress Software Corporation
- Reuters, Ltd.
- RLG
- RWS Group, LLC
- SAP AG
- Sun Microsystems, Inc.
- Sybase, Inc.
- Unisys Corporation
- Trigeminal Software
- Xerox Corporation

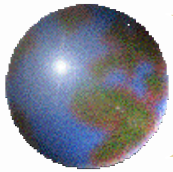
Full Members



The Unicode Consortium®

- Adams Globalization
- Agfa Monotype Corporation
- Beijing Zhong Yi Electronics
- Booz, Allen & Hamilton
- The Church of Jesus Christ of Latter-day Saints
- Columbia University
- DecoType, Inc.
- Endeavor Information Systems
- Ex Libris
- Government of Tamil Nadu
- Innovative Interfaces, Inc.
- Internet Mail Consortium
- Language Analysis Systems
- The Library Corporation
- Netscape Communications
- Nokia
- OCLC, Inc.
- Production First Software
- SAS Institute, Inc.
- Siebel Systems, Inc.
- SIL International
- SIRSI Corporation
- Software AG
- Sony Ericsson
- Symbian, Ltd.
- VTLS, Inc.
- XenCraft
- Yet Another Society

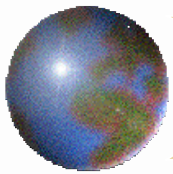
Associate Members



History of Unicode Standard

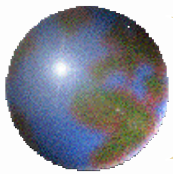
- **Unicode 4.0 (March, 2003)**
- **Unicode 3.2.0 (March, 2002)**
- **Unicode 3.1.1 (August, 2001)**
- **Unicode 3.1.0 (March, 2001)**
- **Unicode 3.0.1 (August, 2000)**
- **Unicode 3.0 (September, 1999)**
- **Unicode 2.0 (July, 1996)**
- **Unicode 1.0 (October, 1991)**

- **The Unicode Consortium. The Unicode Standards.**
  - Version 4.0, 2003. ISBN 0-321-18578-15.
  - Version 3.0, 2000. ISBN 0-201-61633-5.
  - Version 2.0, 1996. ISBN 0-201-48345-9.
  - Version 1.0, Volume 1, 1991. ISBN 0-201-56788-1
    Version 1.0, Volume 2, 1992. ISBN 0-201-60845-6
- **Addison-Wesley Developers Press, Reading, MA.**

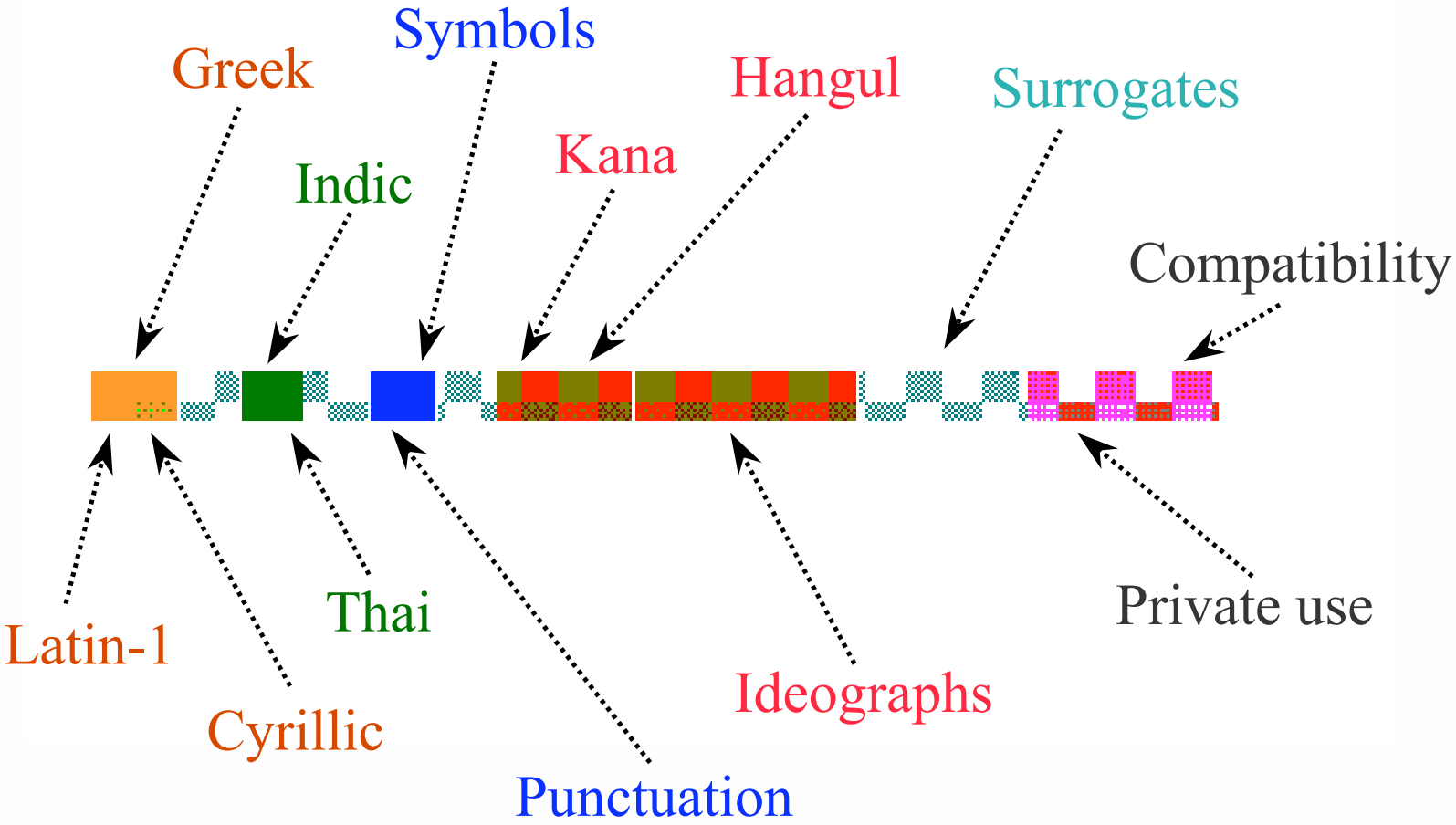


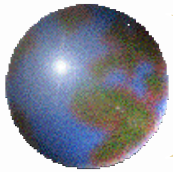
Unicode Principles

- Sixteen bit *representation*
  - Sufficiently compact, but extensible to 16 "planes"
- Characters, not *glyphs*
  - Each abstract character once, except …
- Han, and other, *unification*
  - CJKV hieroglyphics unified when conceptually the same, except …
- Round trip preservation (called "*convertibility*")
  - Hence many a's, alpha, aleph, *etc.*
- Compatibility
  - "wide" characters, Arabic contextual forms, ligatures, *etc.*



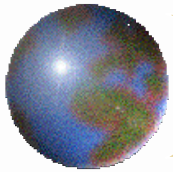
Results





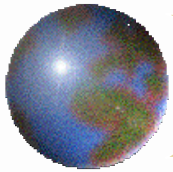
Unicode Character Set

- Arabic
- Armenian
- Bengali
- Bopomofo
- Buhid
- Canadian Syllabics
- Cherokee
- Cypriot
- Cyrillic
- Deseret
- Devanagari
- Ethiopic
- Georgian
- Gothic
- Greek
- Gujarati
- Gurmukhi
- Han
- Hangul
- Hanunóo
- Hebrew
- Hiragana
- Kannada
- Katakana
- Khmer
- Lao
- Latin
- Limbu
- Linear B
- Malayalam
- Mongolian
- Myanmar
- Ogham
- Old Italic (Etruscan)
- Osmanya
- Oriya
- Runic
- Shavian
- Sinhala
- Syriac
- Tagalog
- Tagbanwa
- Tai Le
- Tamil
- Telugu
- Thaana
- Thai
- Tibetan
- Ugaritic
- Yi
- Numbers
- General Diacritics
- General Punctuation
- General Symbols
- Mathematical Symbols
- Musical Symbols (Western & Byzantine)
- Technical Symbols
- Dingbats
- Arrows, Blocks, Box Drawing Forms, and Geometric Shapes
- Miscellaneous Symbols
- Presentation Forms
- Braille Patterns
- Kangxi Radicals



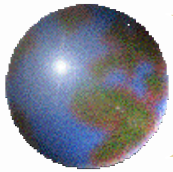
Unicode Truths

- ASCII is in twice ($0021\text{-}007E_{16}$, $FF01\text{-}FF5E_{16}$)
- 30 sets of decimal digits, 0-9
- 17 space characters (not counting tabs, *etc.*)
- 16 hyphen or dash characters
- composed and decomposed characters



Han Unification

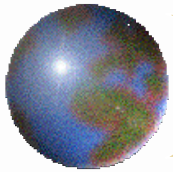
Unicode	China	Taiwan	Japan	Korea
4E00	一	一	一	一
4E0E	与	与	与	
5224	判	判	判	判
5668	器	器	器	器
5B57	字	字	字	字
6D77	海	海	海	海
9038	逸	逸	逸	逸
9AA8	骨	骨	骨	骨



Unicode Characteristics

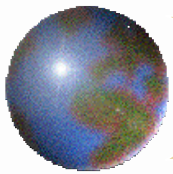
- Character name
- General Category
- Canonical Combining Classes
- Bi-directional Category
- Character Decomposition Mapping
- Decimal digit value
- Digit value
- Numeric value
- Mirrored
- Unicode 1.0 Name
- 10646 comment field
- Upper case Mapping
- Lower case Mapping
- Title case Mapping

`UnicodeData.txt`



Example (1) of UnicodeData.txt

0C66;TELUGU DIGIT ZERO;Nd;0;L;;0;0;0;N;;;;;
0C67;TELUGU DIGIT ONE;Nd;0;L;;1;1;1;N;;;;;
0C68;TELUGU DIGIT TWO;Nd;0;L;;2;2;2;N;;;;;
0C69;TELUGU DIGIT THREE;Nd;0;L;;3;3;3;N;;;;;
0C6A;TELUGU DIGIT FOUR;Nd;0;L;;4;4;4;N;;;;;
0C6B;TELUGU DIGIT FIVE;Nd;0;L;;5;5;5;N;;;;;
0C6C;TELUGU DIGIT SIX;Nd;0;L;;6;6;6;N;;;;;
0C6D;TELUGU DIGIT SEVEN;Nd;0;L;;7;7;7;N;;;;;
0C6E;TELUGU DIGIT EIGHT;Nd;0;L;;8;8;8;N;;;;;
0C6F;TELUGU DIGIT NINE;Nd;0;L;;9;9;9;N;;;;;



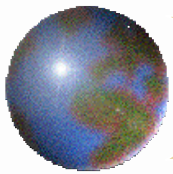
Example (2) of UnicodeData.txt

```
0024;DOLLAR SIGN;Sc;0;ET;;;;;N;;;;;
00A2;CENT SIGN;Sc;0;ET;;;;;N;;;;;
00A3;POUND SIGN;Sc;0;ET;;;;;N;;;;;
00A4;CURRENCY SIGN;Sc;0;ET;;;;;N;;;;;
00A5;YEN SIGN;Sc;0;ET;;;;;N;;;;;
...
20A0;EURO-CURRENCY
    SIGN;Sc;0;ET;;;;;N;;;;;
20A1;COLON SIGN;Sc;0;ET;;;;;N;;;;;
20A2;CRUZEIRO SIGN;Sc;0;ET;;;;;N;;;;;
20A3;FRENCH FRANC SIGN;Sc;0;ET;;;;;N;;;;;
20A4;LIRA SIGN;Sc;0;ET;;;;;N;;;;;
20A5;MILL SIGN;Sc;0;ET;;;;;N;;;;;
20A6;NAIRA SIGN;Sc;0;ET;;;;;N;;;;;
20A7;PESETA SIGN;Sc;0;ET;;;;;N;;;;;
...
20AC;EURO SIGN;Sc;0;ET;;;;;N;;;;;
```

```
20AD;KIP SIGN;Sc;0;ET;;;;;N;;;;;
20AE;TUGRIK SIGN;Sc;0;ET;;;;;N;;;;;
20AF;DRACHMA SIGN;Sc;0;ET;;;;;N;;;;;
20B0;GERMAN PENNY
    SIGN;Sc;0;ET;;;;;N;;;;;
20B1;PESO SIGN;Sc;0;ET;;;;;N;;;;;
FDFC;RIAL SIGN;Sc;0;AL;<isolated>
    0631 06CC 0627 0644;;;;N;;;;;
FE69;SMALL DOLLAR
    SIGN;Sc;0;ET;<small> 0024;;;;N;;;;;
FF04;FULLWIDTH DOLLAR
    SIGN;Sc;0;ET;<wide> 0024;;;;N;;;;;
FFE0;FULLWIDTH CENT
    SIGN;Sc;0;ET;<wide> 00A2;;;;N;;;;;
FFE1;FULLWIDTH POUND
    SIGN;Sc;0;ET;<wide> 00A3;;;;N;;;;;
FFE5;FULLWIDTH YEN
    SIGN;Sc;0;ET;<wide> 00A5;;;;N;;;;;
...
```

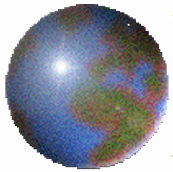
Natural Values and Surrogates

- Natural values 00000-10FFFF$_{16}$
  - Subset of UCS
- Values 10000-10FFFF$_{16}$
  - Planes 1 to 16
  - coded by subtracting 10000$_{16}$, then inserting resulting 20 bits into:
- High: **110110*nnnnnnnnnn*$_2$** (*i.e.*, D800-DBFF$_{16}$)
- Low: **110111*nnnnnnnnnn*$_2$** (*i.e.*, DC00-DFFF$_{16}$)



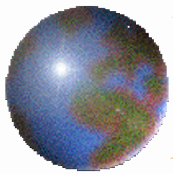
Composed and Decomposed Characters

- Composed: *e.g.*, å è î õ ü (å ≡ U+00E5)
- Decomposed: *e.g.*, a˚ (≡ U+0061 U+030A)
- Multiple accents: *e.g.*, a̋̃ (≡ U+00E5 U+0334
  or U+0061 U+030A U+0334)



Normalization (UAX #15)

- Four forms, the most stringent of which is "compatibility canonical" (KC):
  - all accented characters are decomposed, as well as certain Han decompositions
  - all compatible characters are replaced by their "non-compatible" forms
  - exchangeable accents are exchanged
  - for a given list of composed characters, base characters followed by an accent are recomposed
- *e.g.*, å̃ ("KC" = \u00E5\u0334)



Example (3) of UnicodeData.txt

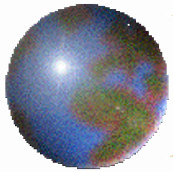
00E4;LATIN SMALL LETTER A WITH DIAERESIS;Ll;0;L;0061 0308;;;;N;LATIN SMALL LETTER A DIAERESIS;;00C4;;00C4

00E5;LATIN SMALL LETTER A WITH RING ABOVE;Ll;0;L;0061 030A;;;;N;LATIN SMALL LETTER A RING;;00C5;;00C5

00E6;LATIN SMALL LETTER AE;Ll;0;L;;;;;N;LATIN SMALL LETTER A E;ash *;00C6;;00C6

00E7;LATIN SMALL LETTER C WITH CEDILLA;Ll;0;L;0063 0327;;;;N;LATIN SMALL LETTER C CEDILLA;;00C7;;00C7

00E8;LATIN SMALL LETTER E WITH GRAVE;Ll;0;L;0065 0300;;;;N;LATIN SMALL LETTER E GRAVE;;00C8;;00C8



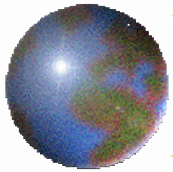
Bidirectionality (UAX #9)

- Mid-Eastern texts inherently bi-directional

0123

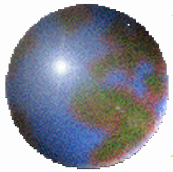
- The Unicode standard (Unicode Standard Annex #9) specifies an embedding algorithm
  - Direction characteristics (strong, weak, neutral)
  - Directionality overrides
  - Language overrides
- The order of characters in a file follows the “natural” order (no directionality).

„« ÂÌ «•‘›—…
«•„ÊÕœ…
"ÌÊ‰ˆfIÊœ" ø



Sorting (UTS #10)

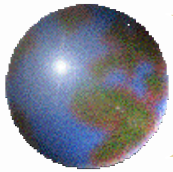
- Sorting by code value does not do the job!
- Unicode specifies four “levels” (KD normalization)
  - *e.g.*, base, case, accent, identity
- Orderings
  - dictionary, language specific
  - telephone directory
  - radical and stroke order, or phonetic, for the Han characters
  - *etc.*



Regular Expressions (UTR #18)

- Ranges specified by:
  - hex codes,
  - Type (digit, letter, separator, *etc.*)
  - Language block (Latin, Greek, Thai, *etc.*)
  - Function (SOL, EOL, white space, *etc.*)
- Other features
  - Level (see Sorting)
  - Normalized/Un-normalized

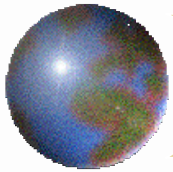
regex



UTF

(Universal Transformation Format)

- UTF-8
  - Transforms UCS-4 into a stream of 8-bit units.
- UTF-16
  - Is the *standard* representation of Unicode in 16-bit units (*i.e.*, with surrogates)
- UTF-32
  - Is the *natural* representation of Unicode in 32-bit units
  - $0\text{-}10FFF_{16}$ (subset of UCS-4).



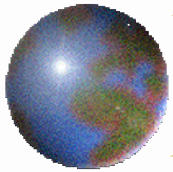
UTF-8

(Transformation of UCS)

- 0nnnnnnn

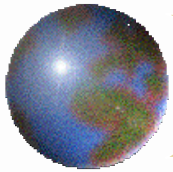
  bits = 7

- 110nnnnn 10nnnnnn
  11
- 1110nnnn 10nnnnnn 10nnnnnn
  16
- 11110nnn 10nnnnnn 10nnnnnn 10nnnnnn
  21
- 111110nn 10nnnnnn 10nnnnnn 10nnnnnn 10nnnnnn
  26



Byte Order Marker

- U+FEFF **ZERO WIDTH NO-BREAK SPACE**
- U+FFFE *not a character code*
- Bytes at the beginning of a file:
  - **FE**$_{16}$ **FF**$_{16}$ *UTF-16 high byte first*
  - **FF**$_{16}$ **EE**$_{16}$ *UTF-16 low byte first*
  - **EF**$_{16}$ **BB**$_{16}$ **BF**$_{16}$ *UTF-8*
  - **00**$_{16}$ **00**$_{16}$ **FE**$_{16}$ **FF**$_{16}$ *UTF-32 high byte first*
  - **00**$_{16}$ **00**$_{16}$ **FF**$_{16}$ **EE**$_{16}$ *UTF-32 low byte first*



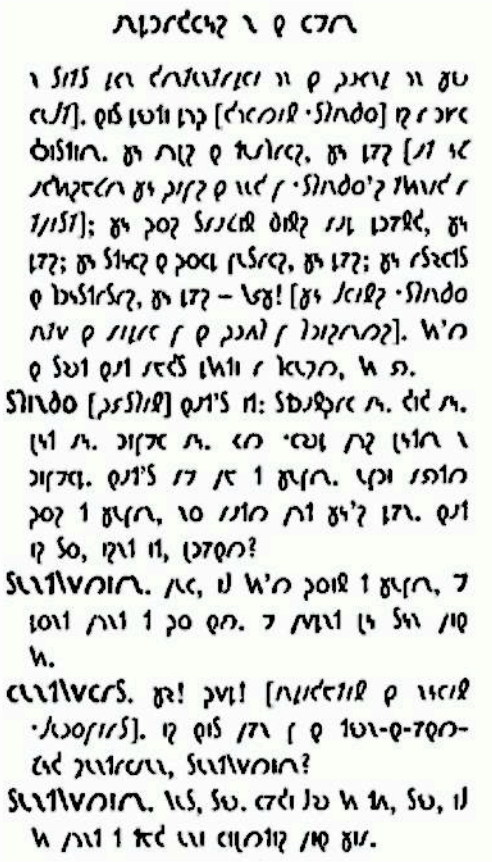
New Scripts in Version 4.0

- BMP, Plane 0
  - Limbu
  - Tai Le
- Plane 1
  - Shavian
  - Linear B
  - Ugaritic Cuneiform
  - Cypriot syllabary
  - Osmanya
- High Voltage Sign ($26A1_{16}$)
- Rejected for 4.0
  - Klingon

Current Statistics

Graphic	**96,248**
Format	**134**
Control	**65**
Private Use	**37,468**
Surrogate	**2,048**
Noncharacter	**66**
Reserved	**878,083**

Shavian Script



ANDROCLES AND THE LION

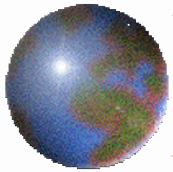
and sits down contentedly on the ground on her left]. This dirty dog [*collaring Spintho*] is a real Christian. He mobs the temples, he does [*at each accusation he gives the neck of Spintho's tunic a twist*]; he goes smashing things mad drunk, he does; he steals the gold vessels, he does; he assaults the priestesses, he does – yah! [*He flings Spintho into the middle of the group of prisoners*]. Youre the sort that makes duty a pleasure, you are.

SPINTHO [*gasping*] Thats it: strangle me. Kick me. Beat me. Revile me. Our Lord was beaten and reviled. Thats my way to heaven. Every martyr goes to heaven, no matter what he's done. That is so, isnt it, brother?

CENTURION. Well, if youre going to heaven, *I* dont want to go there. I wouldnt be seen with you.

LENTULUS. Haw! Good! [*Indicating the kneeling Ferrovius*]. Is this one of the turn-the-other-cheek gentlemen, Centurion?

CENTURION. Yes, sir. Lucky for you too, sir, if you want to take any liberties with him.

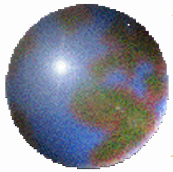


Unicode Applications

- HTML
- XML
- Windows NT/2000, NE (95/98/Me)
- MAC X
- Java
- C/C++ (wchar_t)
- JavaScript
- Browsers (Netscape 4+, IE 5+)
- VB
- Solaris 8, 9 (UTF-8)
- Perl 5.6 (UTF-8)
- Oracle (UTF-8)
- TCL 8.1 (UTF-8)
- Mac 9.0 (UTF-8)

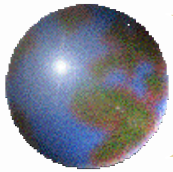
Proposals for New Scripts

- Accepted
  - Glagolitic (BMP)
  - Coptic (BMP)
  - New Testament punctuation (BMP)
  - Ancient Greek acrophonic numerals
  - Ancient Greek papyrological numerals
  - Old Persian
  - Kharoshthi
  - Ancient Greek Musical Notation
  - …
- Proposed
  - Basic Egyptian Hieroglyphics;
  - Cirth
  - Tengwar
  - Meroitic
  - Pollard
  - Blissymbolics
  - …



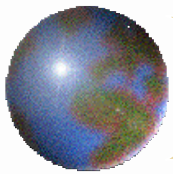
Research Areas

- Sort Specification Languages
- Sort implementation methodologies
- “Large” font management
- Converting to a 21 bit world
- Normalization libraries
- Han refinement
- Archeological research
- Extension of Unicode to further scripts



Summary

- Unicode is more than just another character set, or encoding
- Unicode is “multi-byte, complex”
- Calls into question many of the basic assumptions we make about Western languages
- Has gained much deeper acceptance and understanding in text applications (but still not fully understood)



Further information

- Unicode
  - **http://www.unicode.org/**
- HTML, XML, the Web
  - **http://www.w3.org/TR/unicode-xml/**
    Unicode in XML and other Markup Languages (Unicode Technical Report #20 W3C Note 13 June 2003)
  - **http://www.w3.org/TR/charmod/**
    Character Model for the World Wide Web 1.0 (W3C Working Draft 22 August 2003)
- History
  - **http://www.loc.gov/marc/specifications/speccharucs.html**
- Support
  - **http://java.sun.com/j2se/1.4.2/docs/guide/intl/**
  - **http://oss.software.ibm.com/icu/**
  - **http://search.microsoft.com/search/results.aspx?qu=unicode**