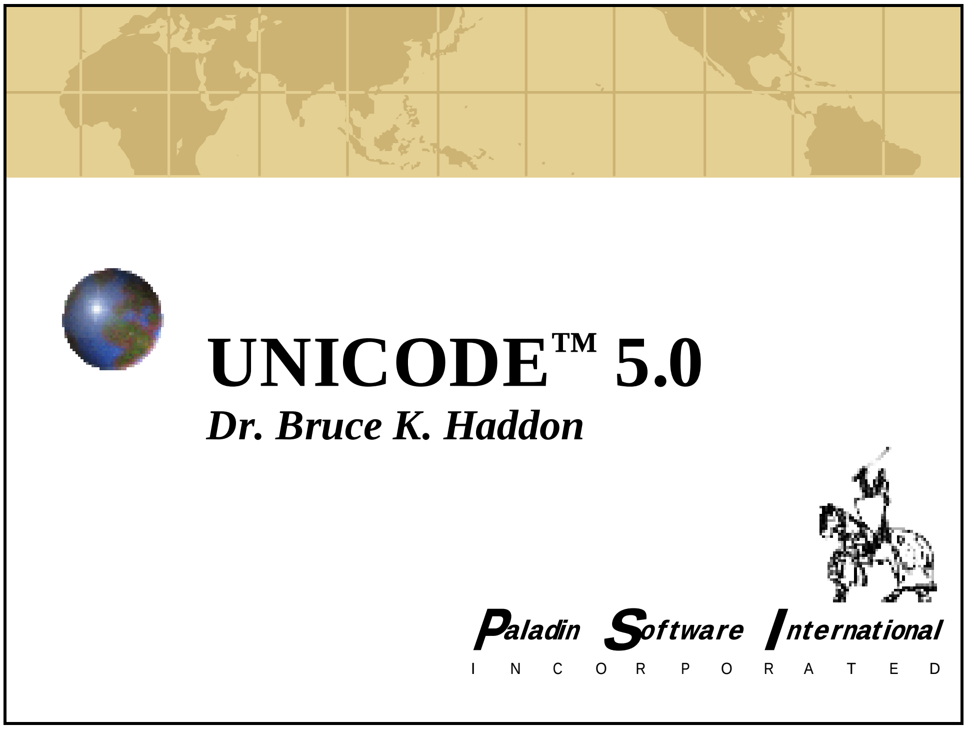
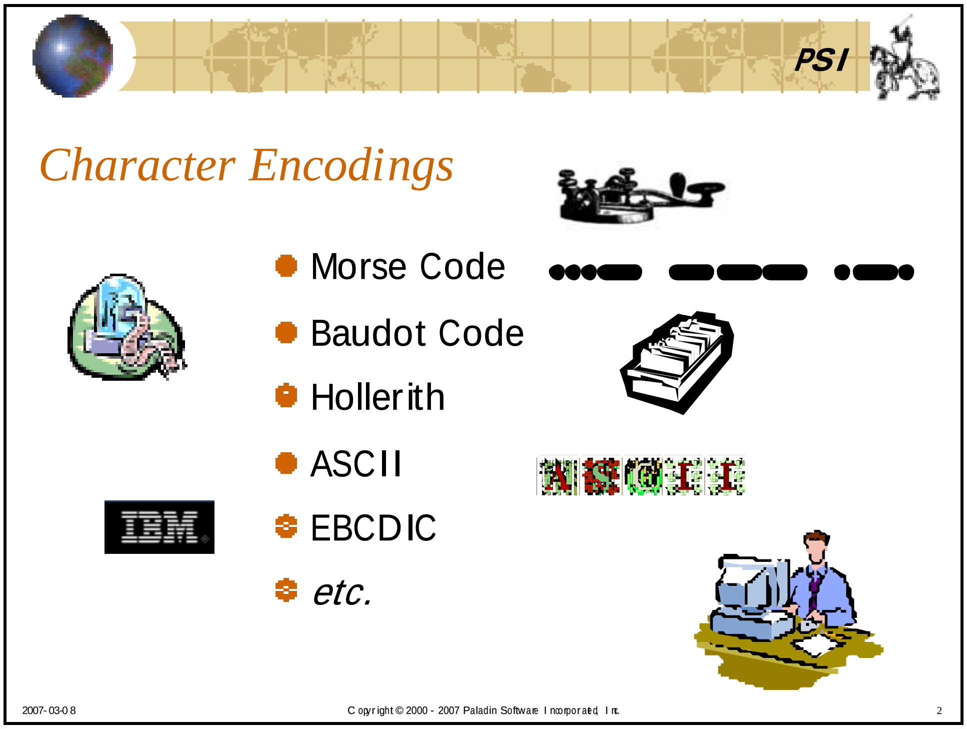


UNICODE™ 5.0

Dr. Bruce K. Haddon

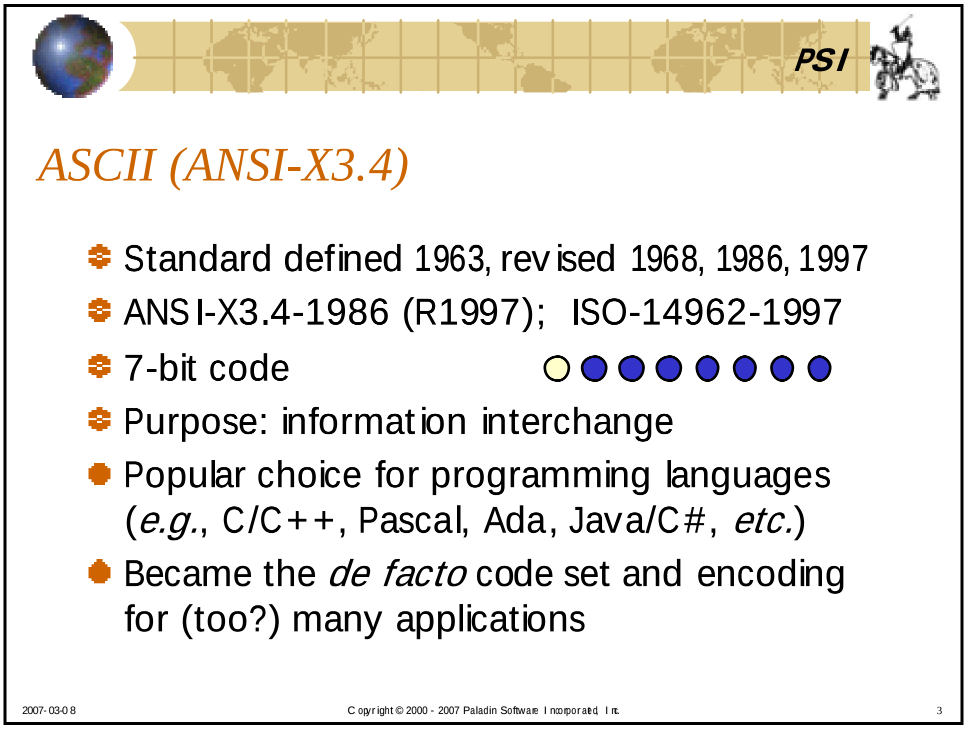
Paladin Software International
INCORPORATED



Character Encodings

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

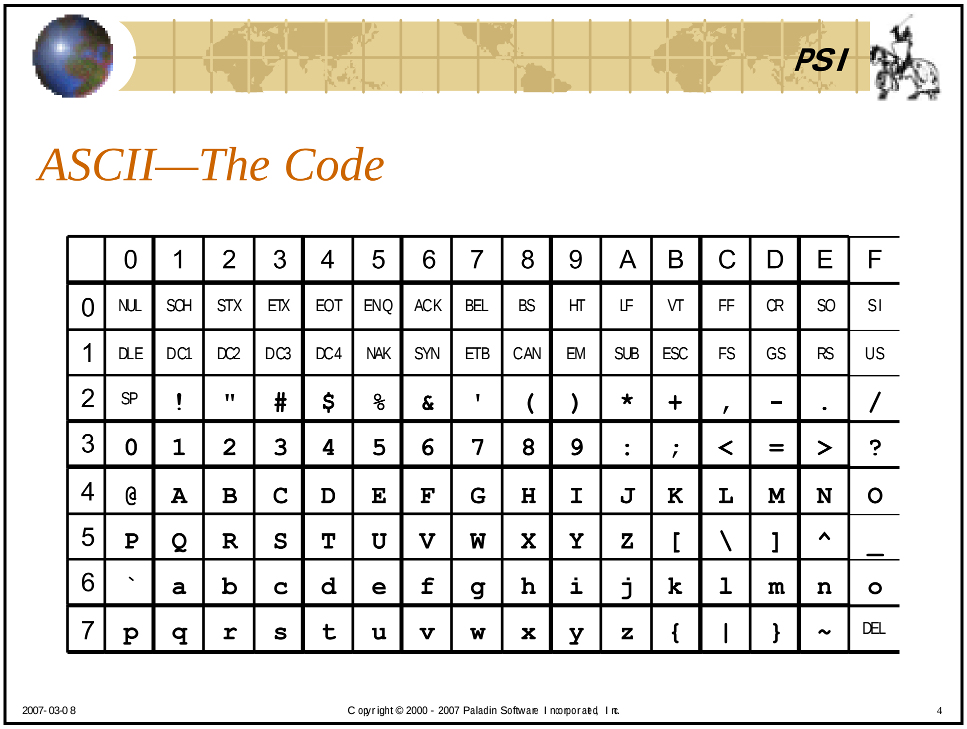
2007-03-08 Copyright © 2000 - 2007 Paladin Software Incorporated, Inc. 2



ASCII (ANSI-X3.4)

- Standard defined 1963, revised 1968, 1986, 1997
- ANSI-X3.4-1986 (R1997); ISO-14962-1997
- 7-bit code ○●●●●●●●
- Purpose: information interchange
- Popular choice for programming languages (*e.g.*, C/C++, Pascal, Ada, Java/C#, *etc.*)
- Became the *de facto* code set and encoding for (too?) many applications

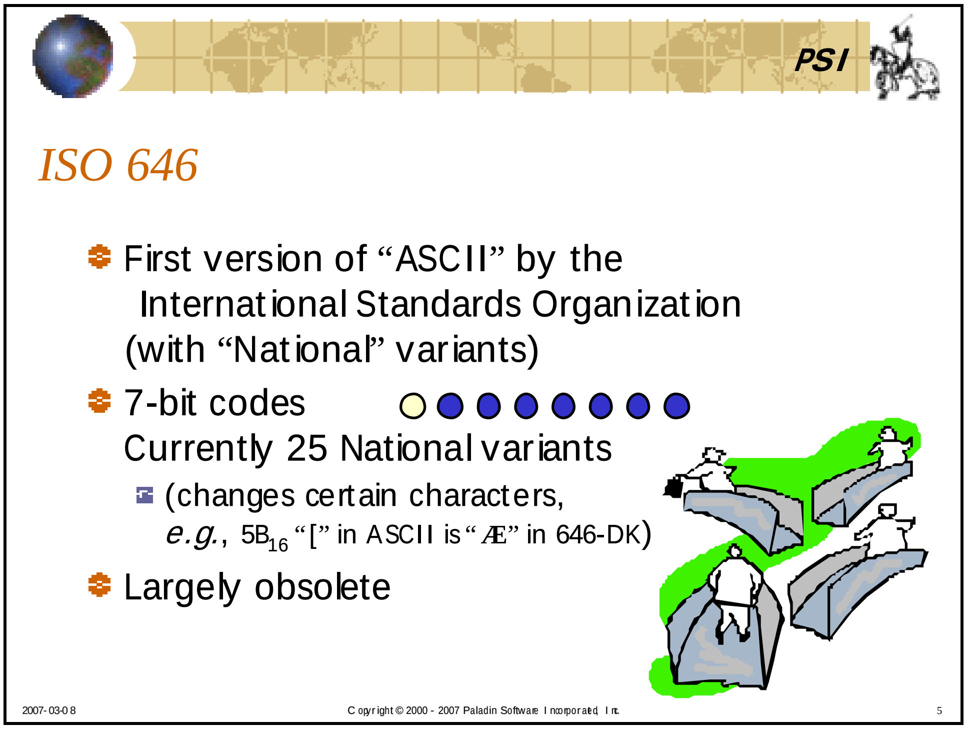
2007-03-08 Copyright © 2000 - 2007 Paladin Software Incorporated, Inc. 3



ASCII—The Code

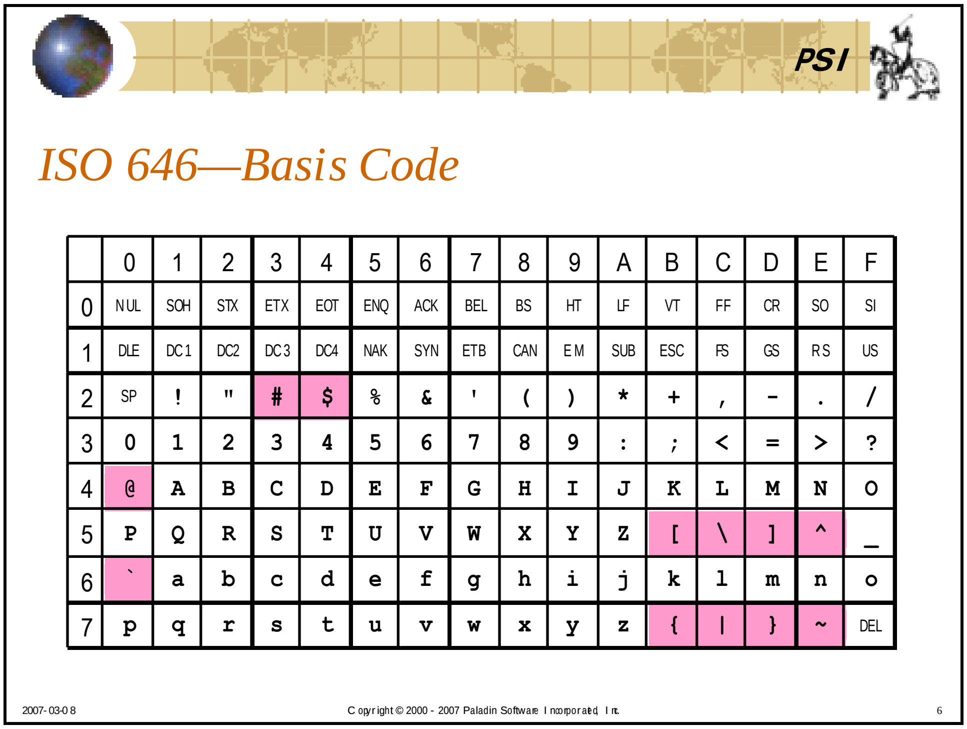
	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SCH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2	SP	!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	e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

2007-03-08 Copyright © 2000 - 2007 Paladin Software Incorporated, Inc. 4



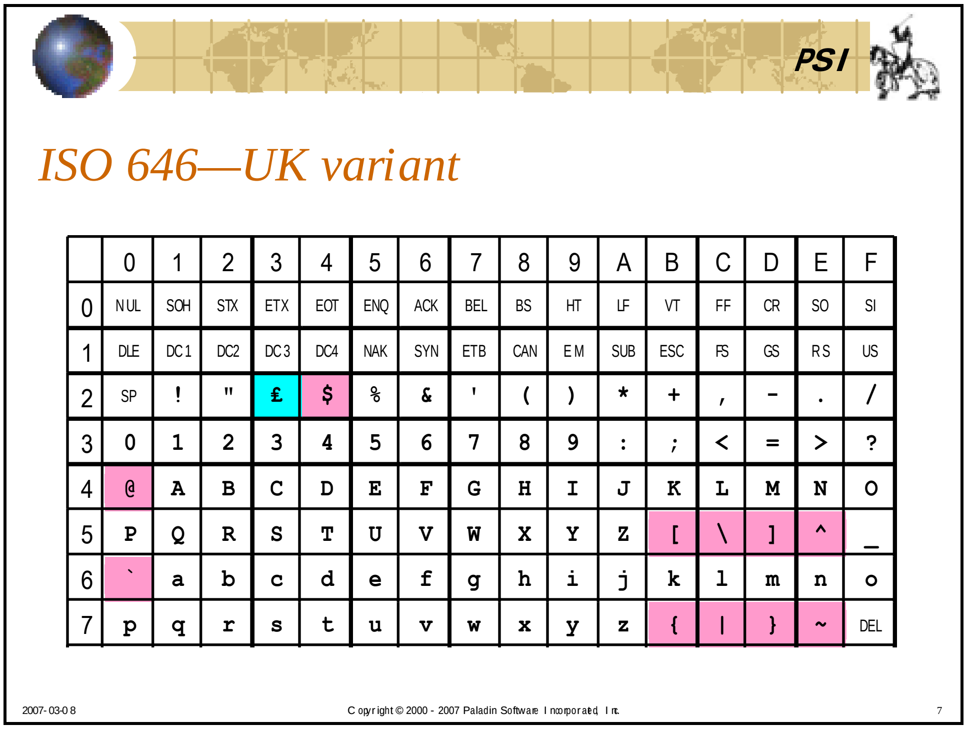
ISO 646

- First version of “ASCII” by the International Standards Organization (with “National” variants)
- 7-bit codes ○●●●●●●●
  Currently 25 National variants
  - (changes certain characters, *e.g.*, $5B_{16}$ “[” in ASCII is “Æ” in 646-DK)
- Largely obsolete



ISO 646—Basis 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

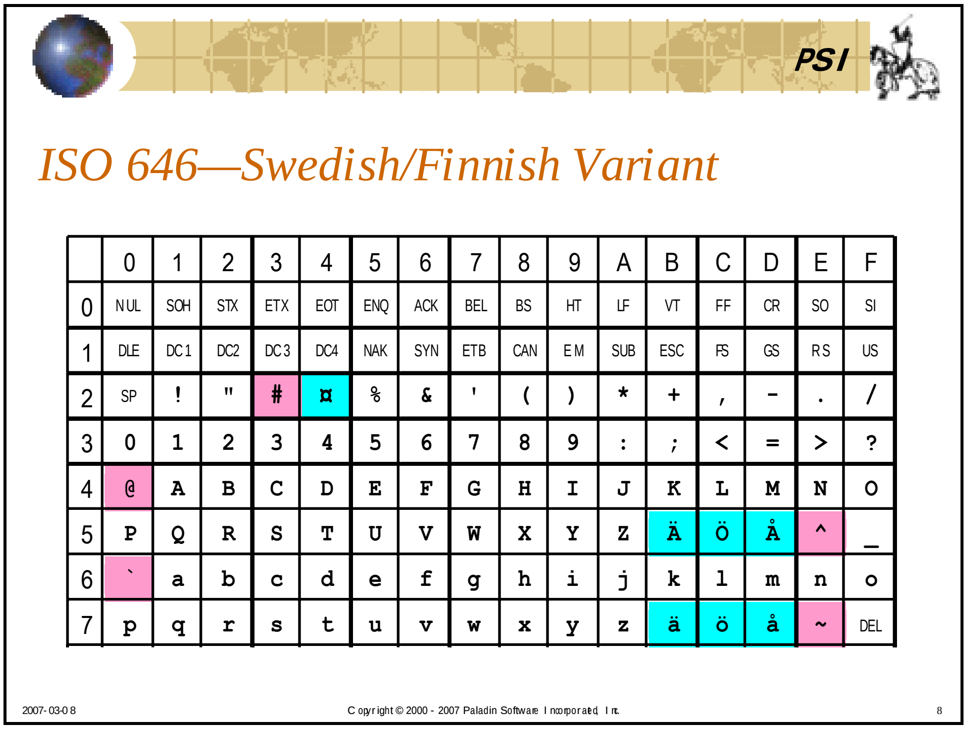


PSI

ISO 646—UK variant

	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

2007-03-08 Copyright © 2000 - 2007 Paladin Software Incorporated, Inc. 7

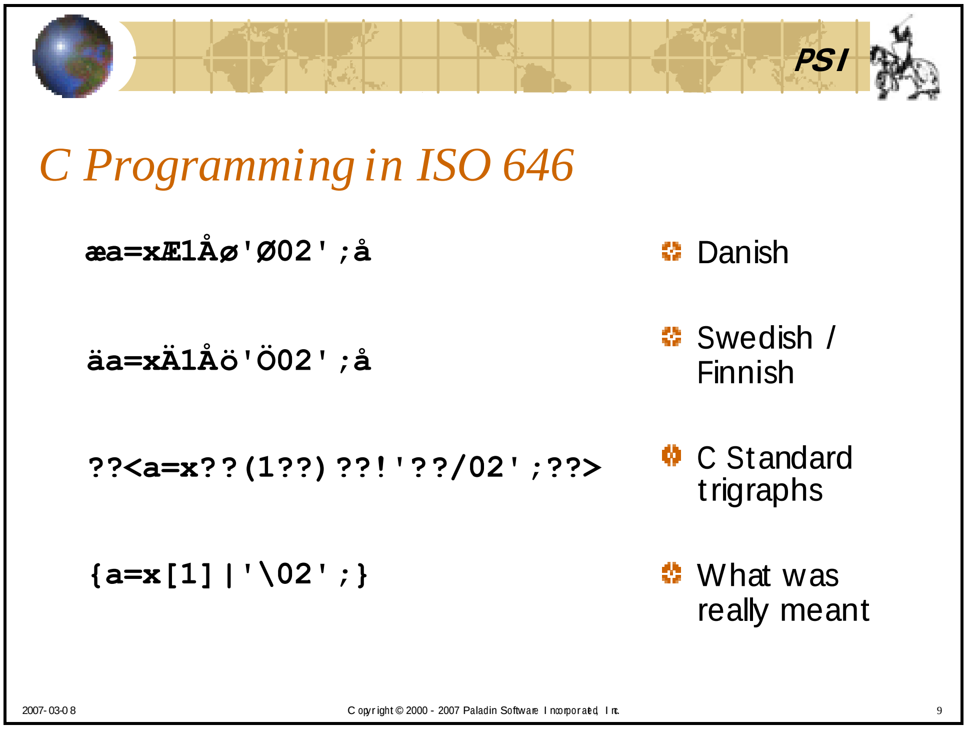


PSI

ISO 646—Swedish/Finnish Variant

	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	Ä	Ö	Å	^	_
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	ä	ö	å	~	DEL

2007-03-08 Copyright © 2000 - 2007 Paladin Software Incorporated, Inc. 8



PSI

C Programming in ISO 646

```
æa=xÆ1Åø'Ø02';å
```

- Danish

```
äa=xÄ1Åö'Ö02';å
```

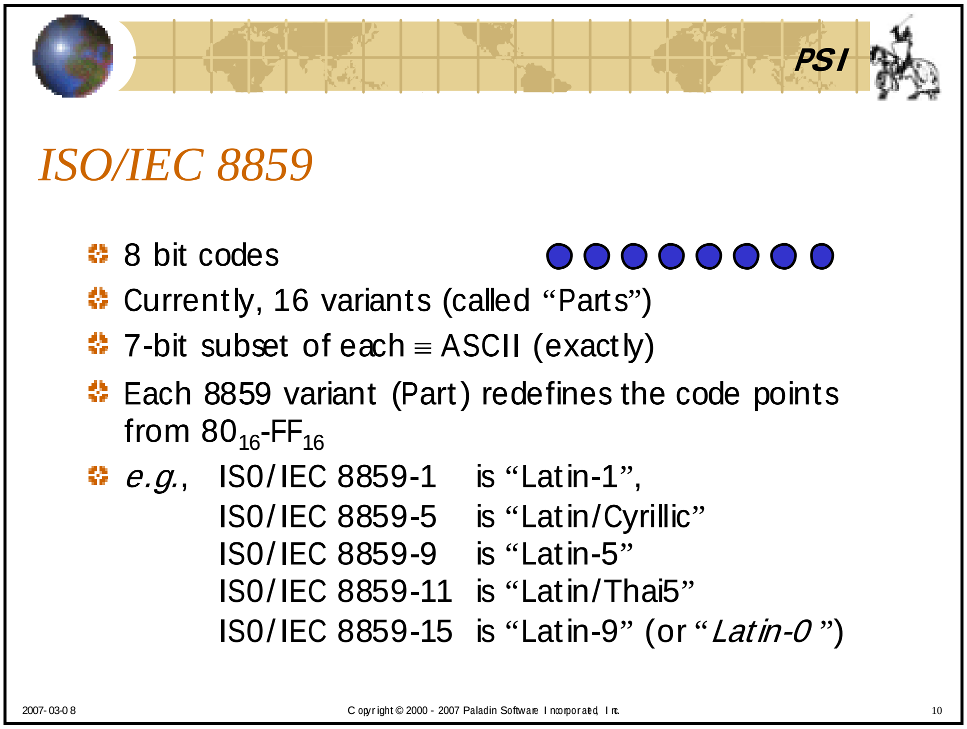
- Swedish / Finnish

```
??<a=x??(1??)??!'??/02';??>
```

- C Standard trigraphs

```
{a=x[1]|'\02';}
```

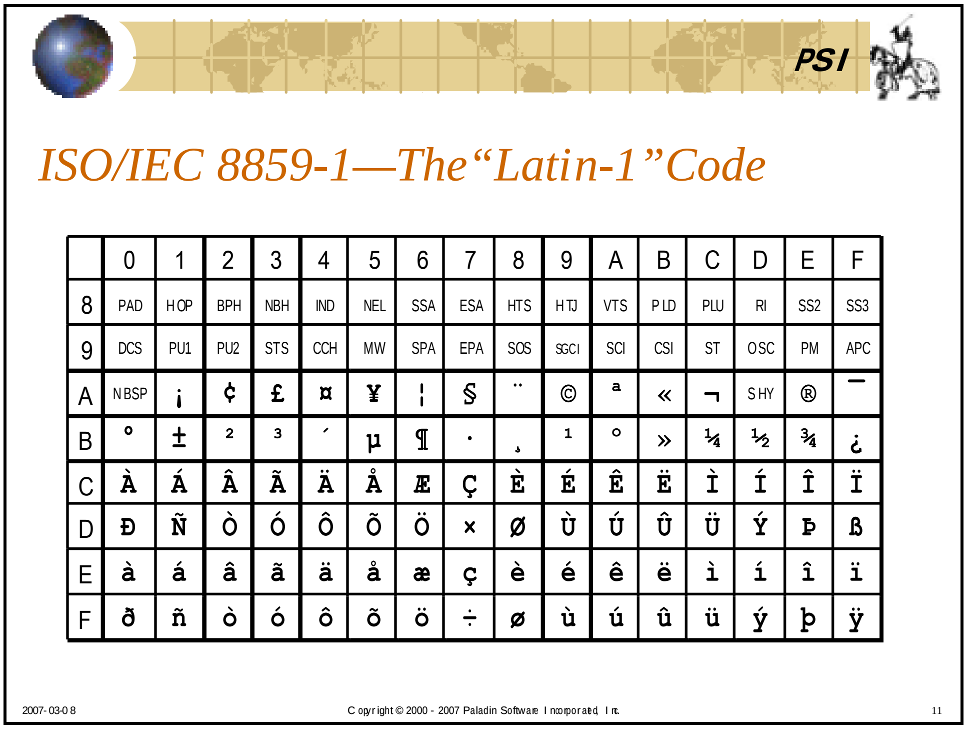
- What was really meant



PSI

ISO/IEC 8859

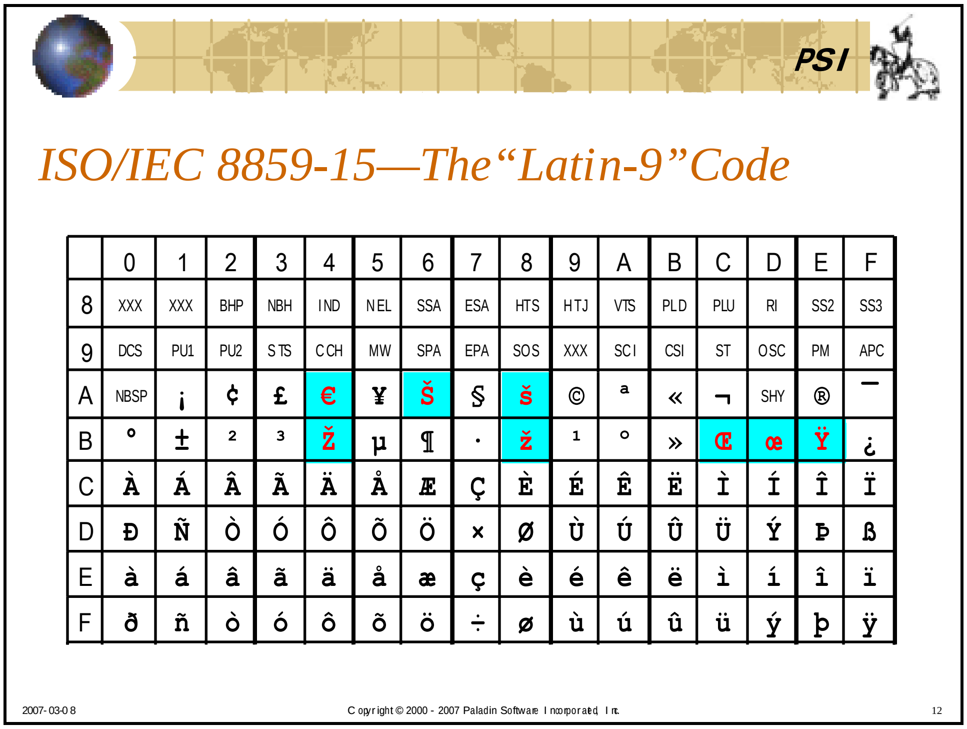
- 8 bit codes ● ● ● ● ● ● ● ●
- Currently, 16 variants (called “Parts”)
- 7-bit subset of each ≡ ASCII (exactly)
- Each 8859 variant (Part) redefines the code points from 80_{16}-FF_{16}
- *e.g.*, ISO/IEC 8859-1 is “Latin-1”,
  ISO/IEC 8859-5 is “Latin/Cyrillic”
  ISO/IEC 8859-9 is “Latin-5”
  ISO/IEC 8859-11 is “Latin/Thai5”
  ISO/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	PAD	HOP	BPH	NBH	IND	NEL	SSA	ESA	HTS	HTJ	VTS	PLD	PLU	RI	SS2	SS3
9	DCS	PU1	PU2	STS	CCH	MW	SPA	EPA	SOS	SGCI	SCI	CSI	ST	OSC	PM	APC
A	NBSP	¡	¢	£	¤	¥	¦	§	¨	©	ª	«	¬	SHY	®	¯
B	°	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿
C	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
D	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	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	ý	þ	ÿ

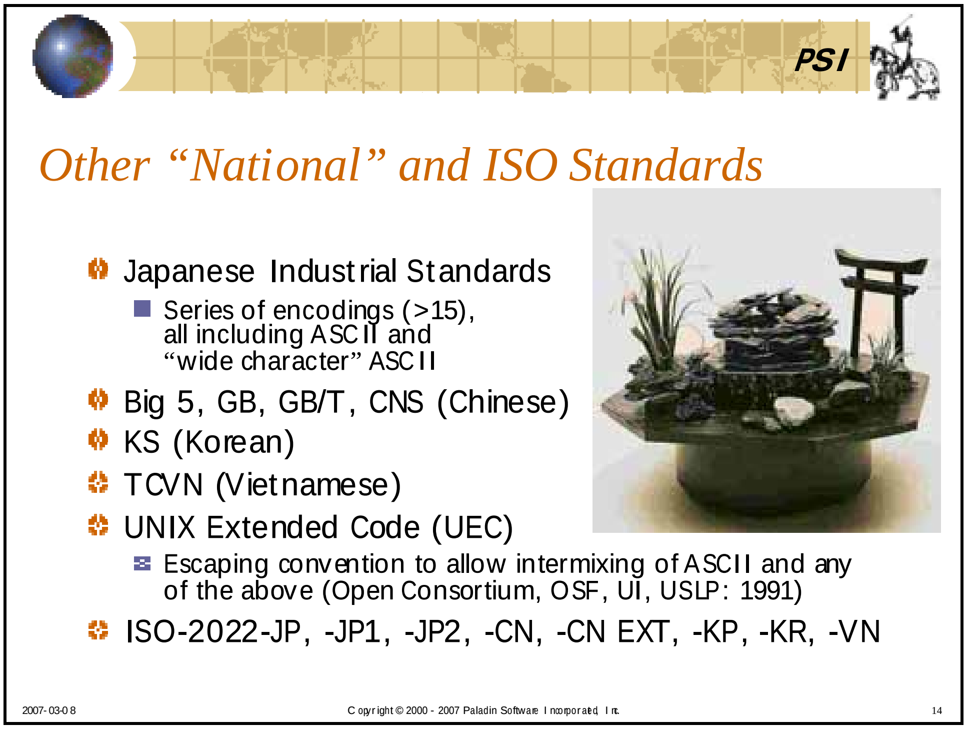
2007-03-08 Copyright © 2000 - 2007 Paladin Software Incorporated, Inc. 11



ISO/IEC 8859-15—The“Latin-9”Code

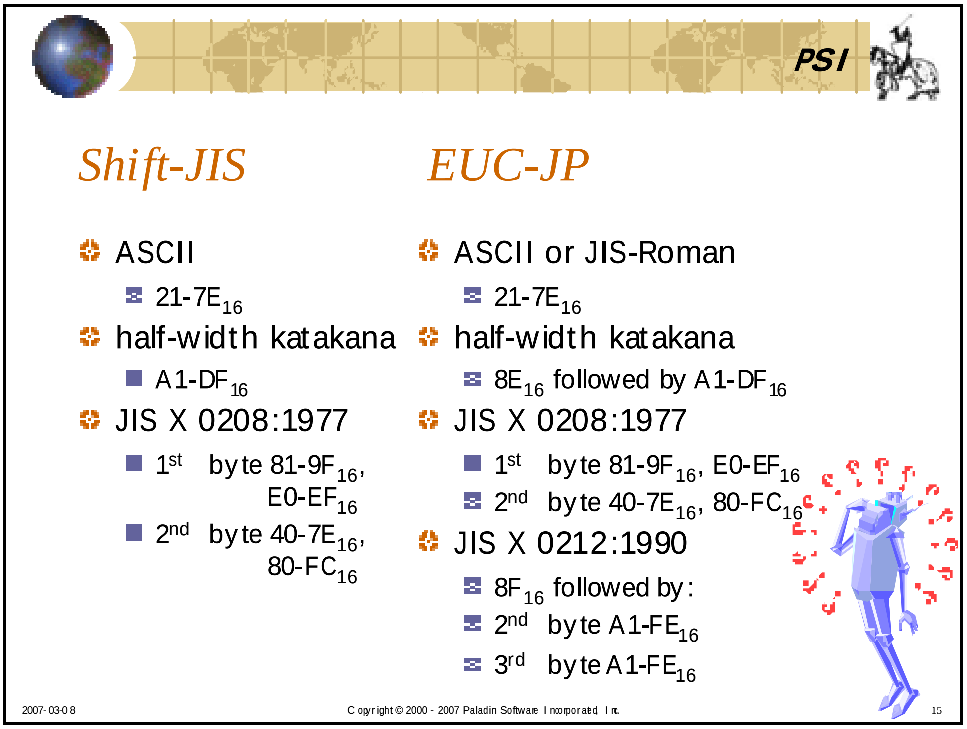
	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
8	XXX	XXX	BHP	NBH	IND	NEL	SSA	ESA	HTS	HTJ	VTS	PLD	PLU	RI	SS2	SS3
9	DCS	PU1	PU2	STS	CCH	MW	SPA	EPA	SOS	XXX	SCI	CSI	ST	OSC	PM	APC
A	NBSP	¡	¢	£	€	¥	Š	§	š	©	ª	«	¬	SHY	®	¯
B	°	±	²	³	Ž	µ	¶	·	ž	¹	º	»	Œ	œ	Ÿ	¿
C	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
D	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
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	ý	þ	ÿ

2007-03-08 Copyright © 2000 - 2007 Paladin Software Incorporated, Inc. 12



Other "National" and ISO Standards

- Japanese Industrial Standards
  - Series of encodings (>15), all including ASCII and "wide character" ASCII
- Big 5, GB, GB/T, CNS (Chinese)
- KS (Korean)
- TCVN (Vietnamese)
- 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

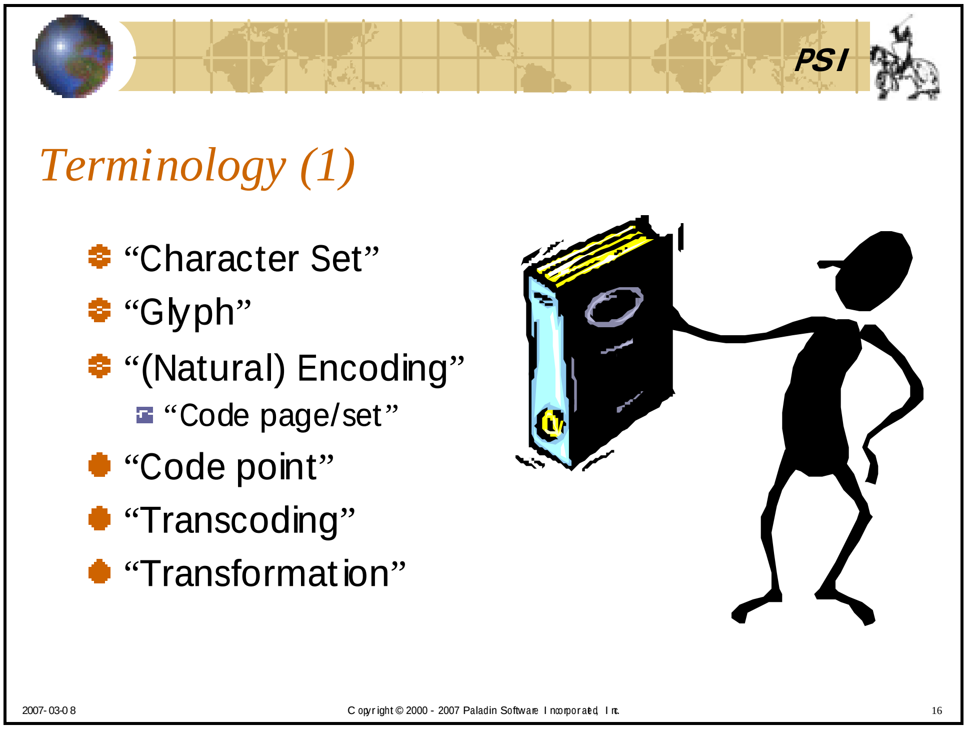


Shift-JIS

- ASCII
  - 21-7E$_{16}$
- half-width katakana
  - A1-DF$_{16}$
- JIS X 0208:1977
  - 1st byte 81-9F$_{16}$, E0-EF$_{16}$
  - 2nd byte 40-7E$_{16}$, 80-FC$_{16}$

EUC-JP

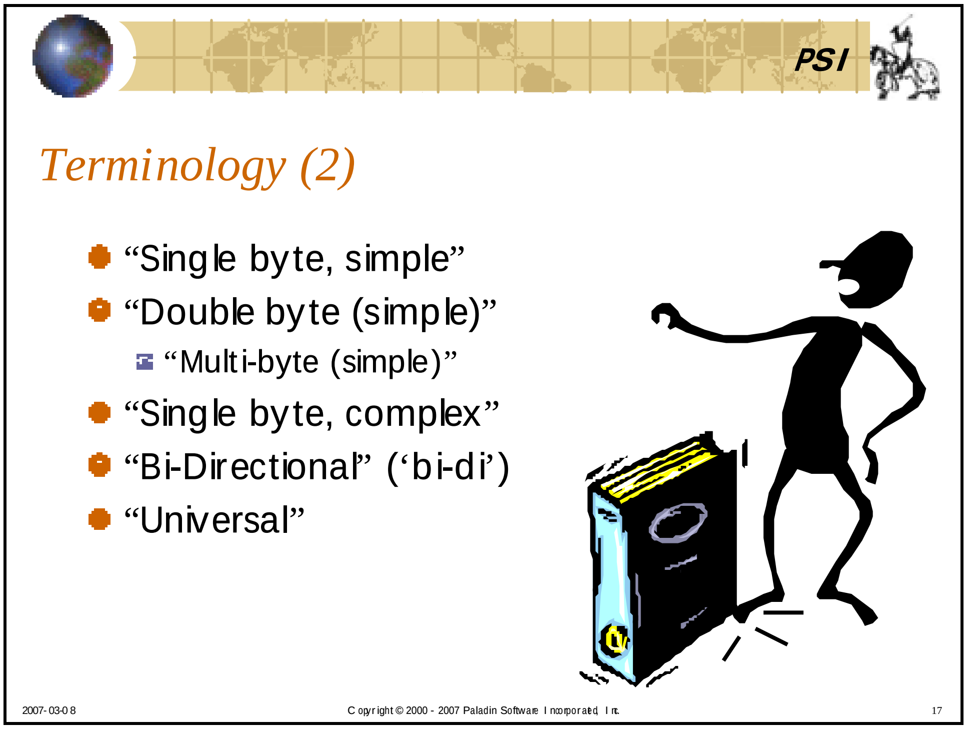
- ASCII or JIS-Roman
  - 21-7E$_{16}$
- half-width katakana
  - 8E$_{16}$ followed by A1-DF$_{16}$
- JIS X 0208:1977
  - 1st byte 81-9F$_{16}$, E0-EF$_{16}$
  - 2nd byte 40-7E$_{16}$, 80-FC$_{16}$
- JIS X 0212:1990
  - 8F$_{16}$ followed by:
  - 2nd byte A1-FE$_{16}$
  - 3rd byte A1-FE$_{16}$



Terminology (1)

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

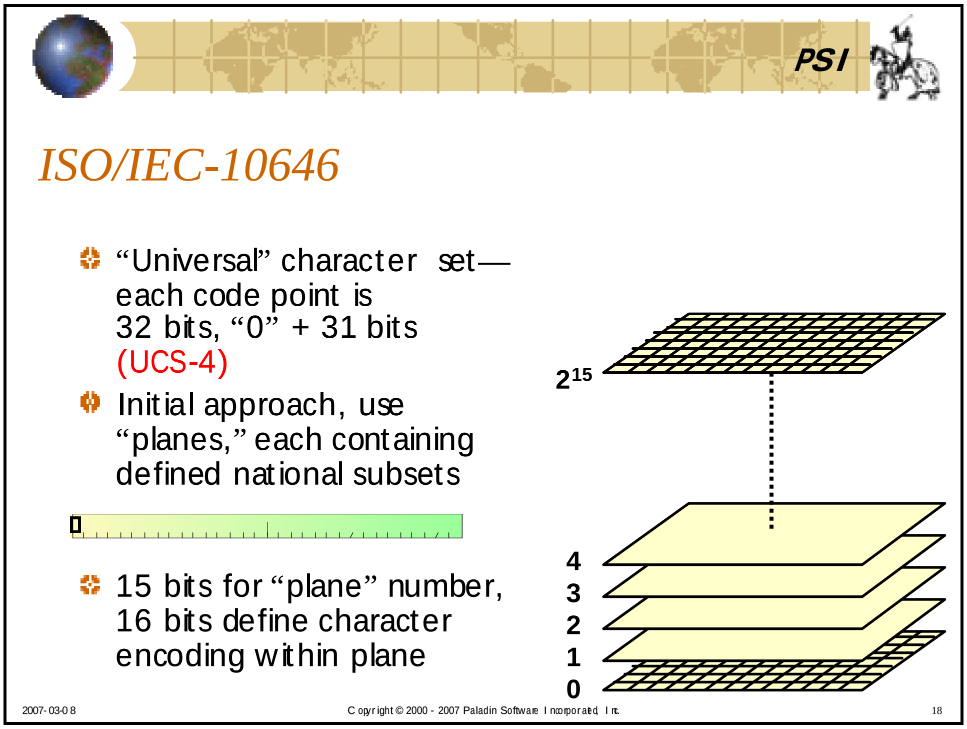
2007-03-08 Copyright © 2000 - 2007 Paladin Software Incorporated, Inc. 16



Terminology (2)

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

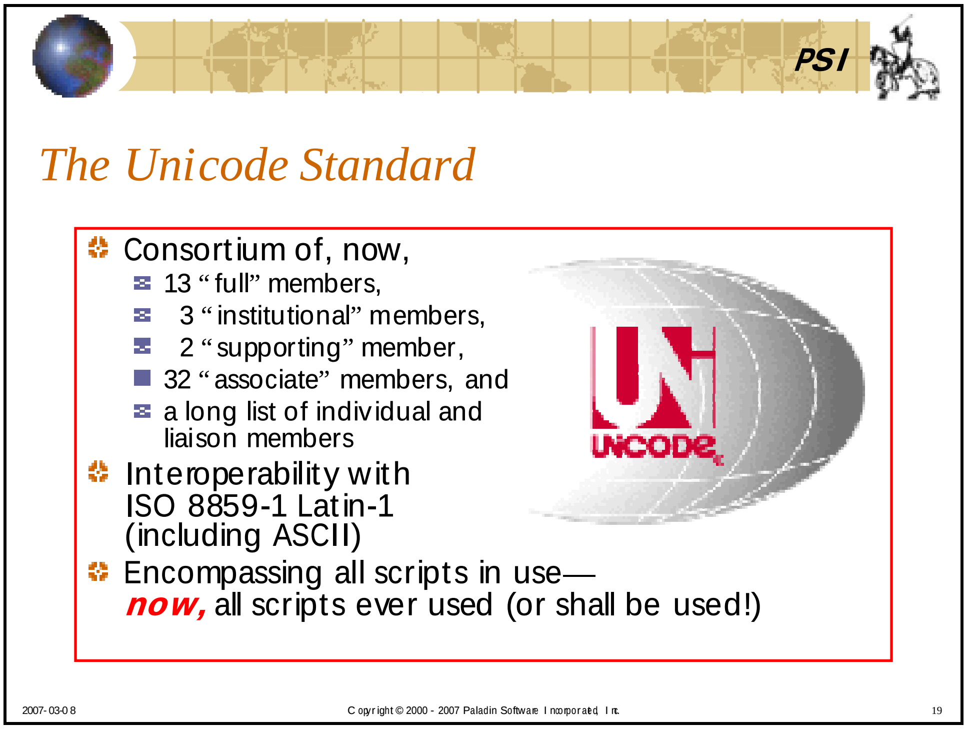
2007-03-08 Copyright © 2000 - 2007 Paladin Software Incorporated, Inc. 17



PSI

ISO/IEC-10646

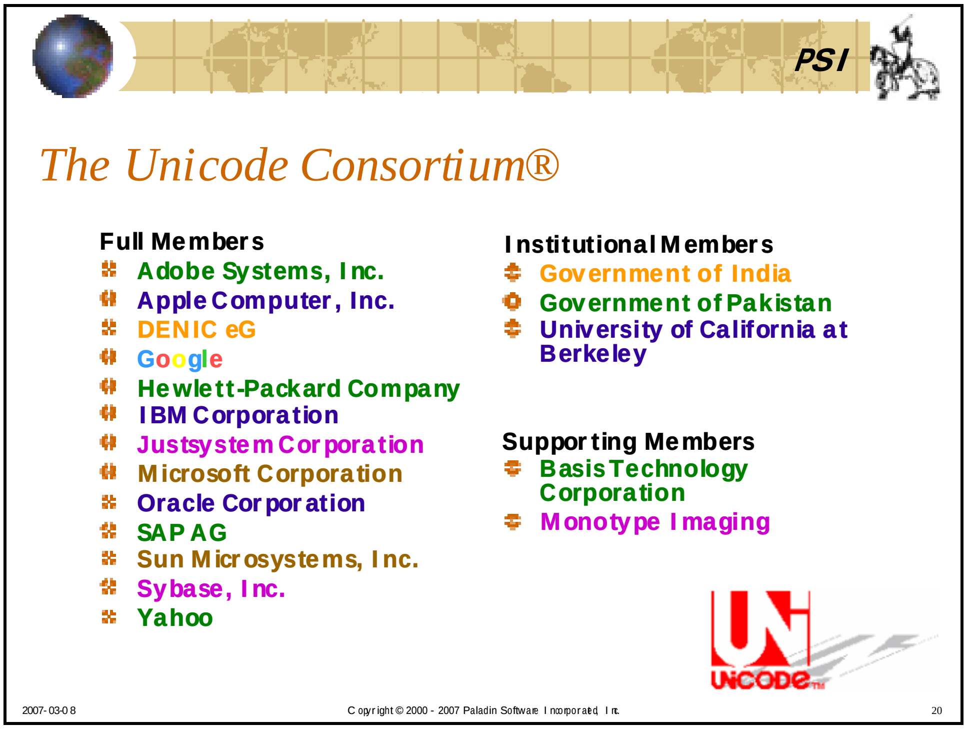
- “Universal” character set—each code point is 32 bits, “0” + 31 bits (UCS-4)
- Initial approach, use “planes,” each containing defined national subsets
- 15 bits for “plane” number, 16 bits define character encoding within plane



PSI

The Unicode Standard

- Consortium of, now,
  - 13 “full” members,
  - 3 “institutional” members,
  - 2 “supporting” member,
  - 32 “associate” members, and
  - a long list of individual and liaison 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®

Full Members

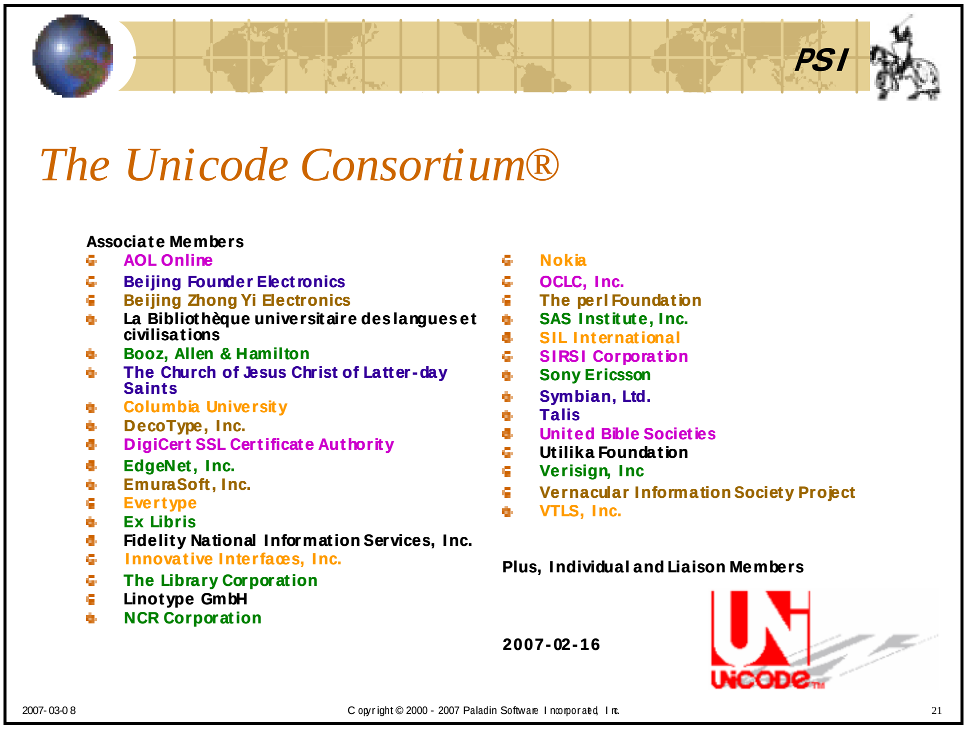
- Adobe Systems, Inc.
- Apple Computer, Inc.
- DENIC eG
- Google
- Hewlett-Packard Company
- IBM Corporation
- Justsystem Corporation
- Microsoft Corporation
- Oracle Corporation
- SAP AG
- Sun Microsystems, Inc.
- Sybase, Inc.
- Yahoo

Institutional Members

- Government of India
- Government of Pakistan
- University of California at Berkeley

Supporting Members

- Basis Technology Corporation
- Monotype Imaging



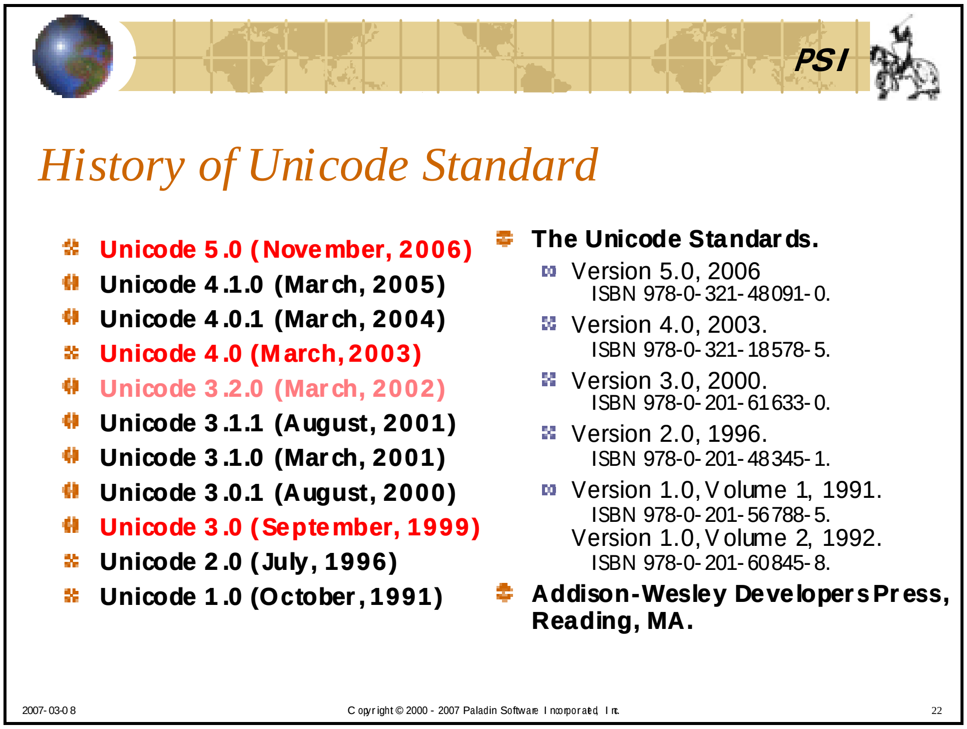
The Unicode Consortium®

Associate Members

- AOL Online
- Beijing Founder Electronics
- Beijing Zhong Yi Electronics
- La Bibliothèque universitaire des langues et civilisations
- Booz, Allen & Hamilton
- The Church of Jesus Christ of Latter-day Saints
- Columbia University
- DecoType, Inc.
- DigiCert SSL Certificate Authority
- EdgeNet, Inc.
- EmuraSoft, Inc.
- Evertype
- Ex Libris
- Fidelity National Information Services, Inc.
- Innovative Interfaces, Inc.
- The Library Corporation
- Linotype GmbH
- NCR Corporation
- Nokia
- OCLC, Inc.
- The perl Foundation
- SAS Institute, Inc.
- SIL International
- SIRSI Corporation
- Sony Ericsson
- Symbian, Ltd.
- Talis
- United Bible Societies
- Utilika Foundation
- Verisign, Inc
- Vernacular Information Society Project
- VTLS, Inc.

Plus, Individual and Liaison Members

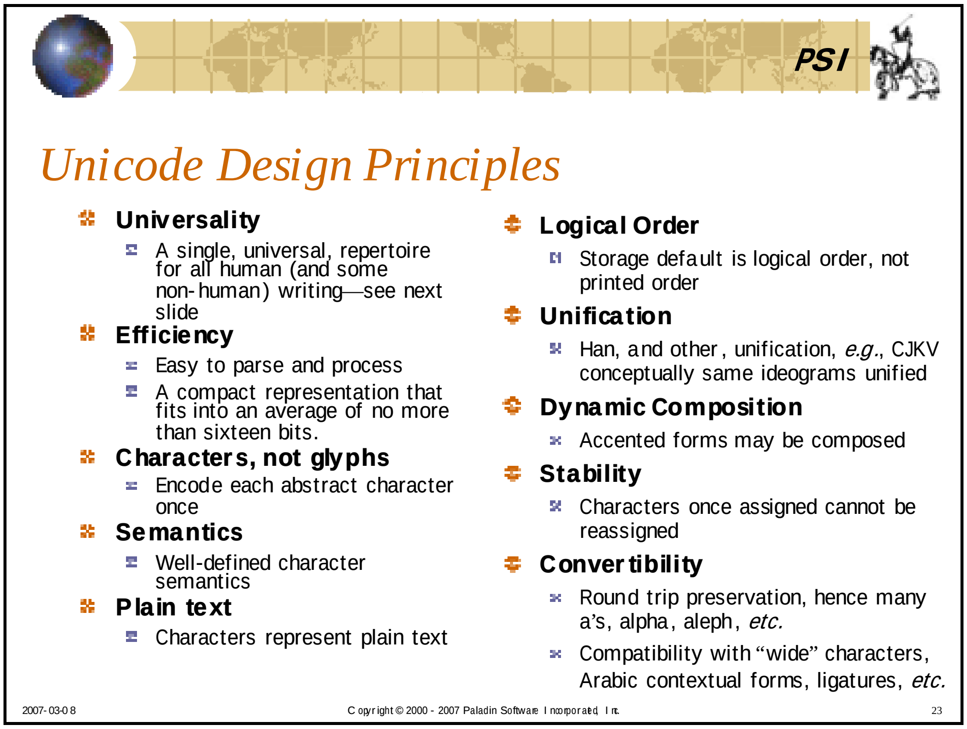
2007-02-16



History of Unicode Standard

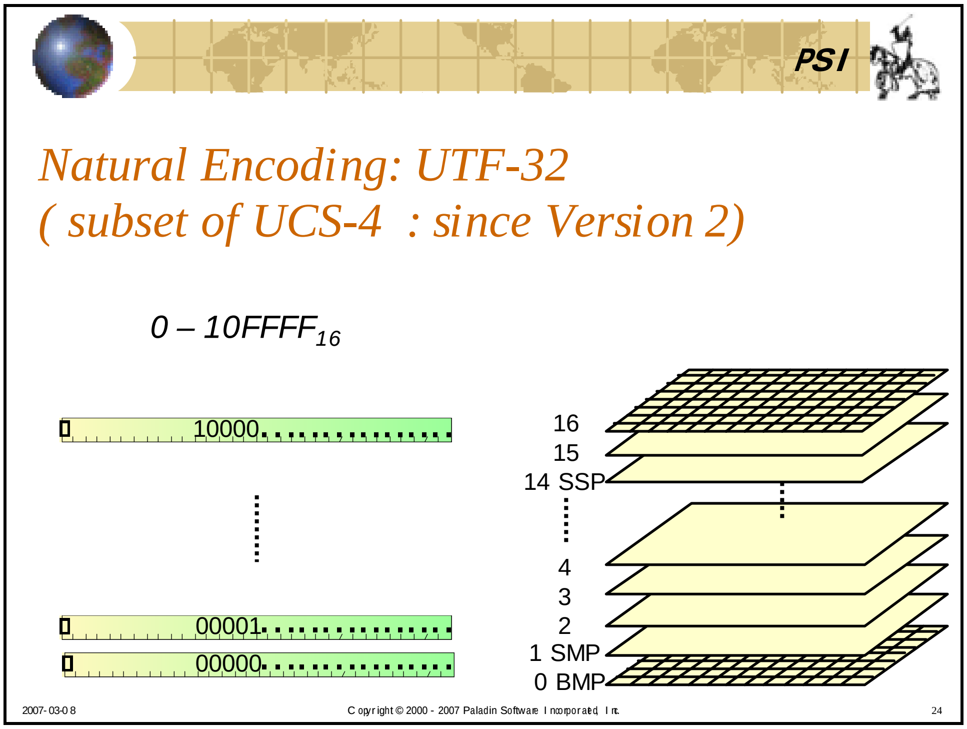
- **Unicode 5.0 (November, 2006)**
- **Unicode 4.1.0 (March, 2005)**
- **Unicode 4.0.1 (March, 2004)**
- **Unicode 4.0 (March, 2003)**
- **Unicode 3.2.0 (March, 2002)**
- **Unicode 3.1.1 (August, 2001)**
- **Unicode 3.1.0 (March, 2001)**
- **Unicode 3.0.1 (August, 2000)**
- **Unicode 3.0 (September, 1999)**
- **Unicode 2.0 (July, 1996)**
- **Unicode 1.0 (October, 1991)**

- **The Unicode Standards.**
  - Version 5.0, 2006
    ISBN 978-0-321-48091-0.
  - Version 4.0, 2003.
    ISBN 978-0-321-18578-5.
  - Version 3.0, 2000.
    ISBN 978-0-201-61633-0.
  - Version 2.0, 1996.
    ISBN 978-0-201-48345-1.
  - Version 1.0, Volume 1, 1991.
    ISBN 978-0-201-56788-5.
    Version 1.0, Volume 2, 1992.
    ISBN 978-0-201-60845-8.
- **Addison-Wesley Developers Press, Reading, MA.**



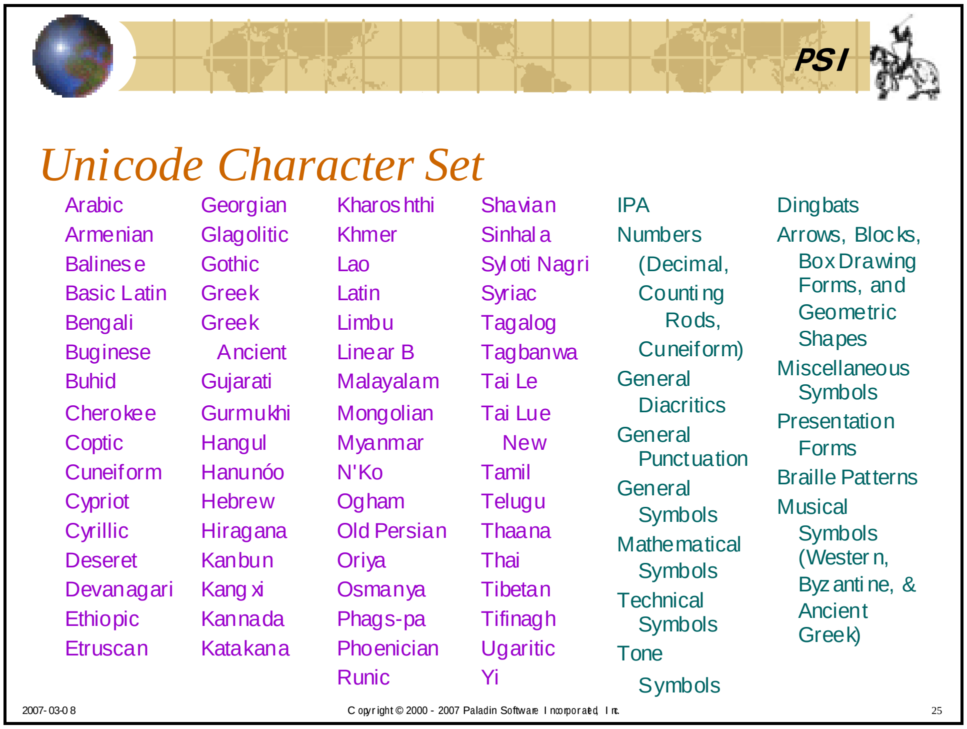
Unicode Design Principles

- **Universality**
  - A single, universal, repertoire for all human (and some non-human) writing—see next slide
- **Efficiency**
  - Easy to parse and process
  - A compact representation that fits into an average of no more than sixteen bits.
- **Characters, not glyphs**
  - Encode each abstract character once
- **Semantics**
  - Well-defined character semantics
- **Plain text**
  - Characters represent plain text
- **Logical Order**
  - Storage default is logical order, not printed order
- **Unification**
  - Han, and other, unification, *e.g.*, CJKV conceptually same ideograms unified
- **Dynamic Composition**
  - Accented forms may be composed
- **Stability**
  - Characters once assigned cannot be reassigned
- **Convertibility**
  - Round trip preservation, hence many a's, alpha, aleph, *etc.*
  - Compatibility with "wide" characters, Arabic contextual forms, ligatures, *etc.*



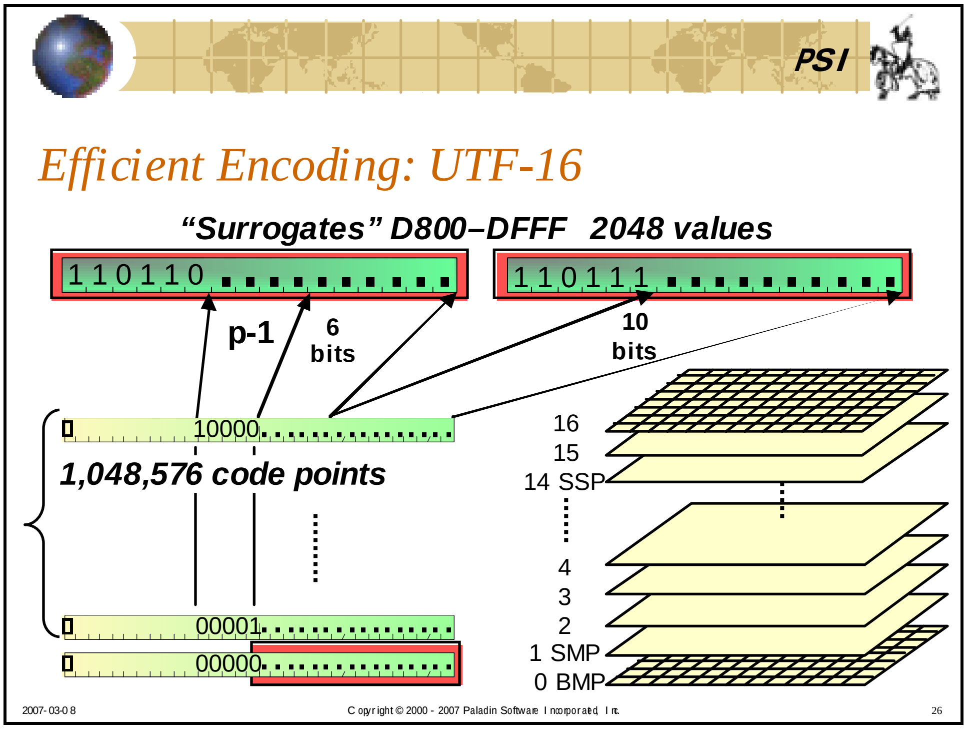
Natural Encoding: UTF-32 (subset of UCS-4 : since Version 2)

0 – 10FFFF$_{16}$



Unicode Character Set

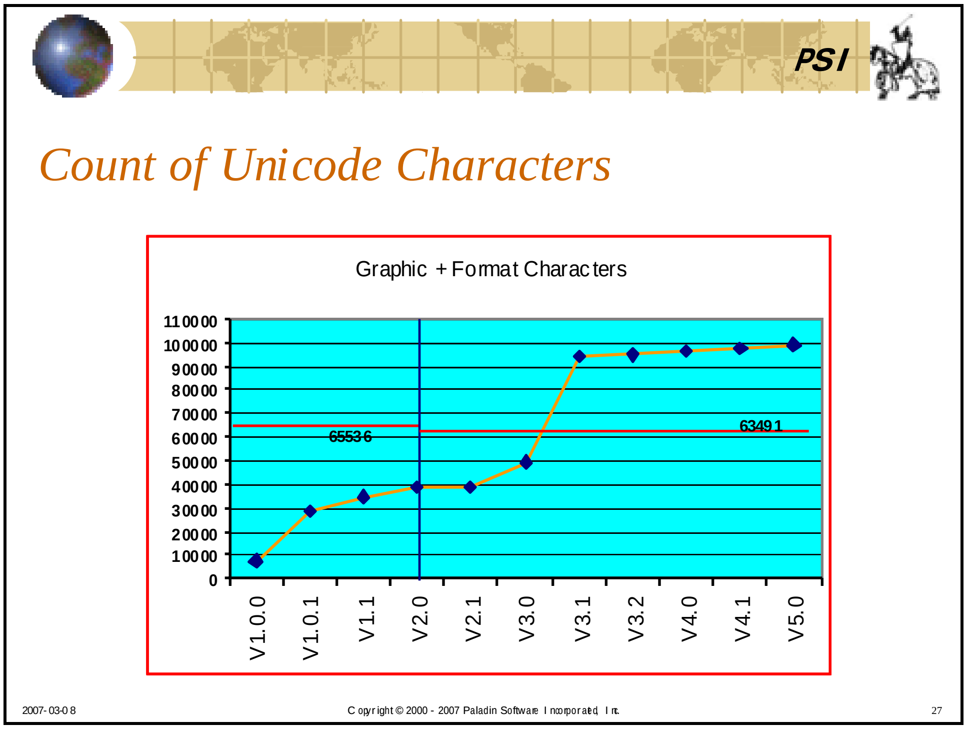
- Arabic
- Armenian
- Balinese
- Basic Latin
- Bengali
- Buginese
- Buhid
- Cherokee
- Coptic
- Cuneiform
- Cypriot
- Cyrillic
- Deseret
- Devanagari
- Ethiopic
- Etruscan
- Georgian
- Glagolitic
- Gothic
- Greek
- Greek Ancient
- Gujarati
- Gurmukhi
- Hangul
- Hanunóo
- Hebrew
- Hiragana
- Kanbun
- Kangxi
- Kannada
- Katakana
- Kharoshthi
- Khmer
- Lao
- Latin
- Limbu
- Linear B
- Malayalam
- Mongolian
- Myanmar
- N'Ko
- Ogham
- Old Persian
- Oriya
- Osmanya
- Phags-pa
- Phoenician
- Runic
- Shavian
- Sinhala
- Syloti Nagri
- Syriac
- Tagalog
- Tagbanwa
- Tai Le
- Tai Lue New
- Tamil
- Telugu
- Thaana
- Thai
- Tibetan
- Tifinagh
- Ugaritic
- Yi
- IPA
- Numbers (Decimal, Counting Rods, Cuneiform)
- General Diacritics
- General Punctuation
- General Symbols
- Mathematical Symbols
- Technical Symbols
- Tone Symbols
- Dingbats
- Arrows, Blocks, Box Drawing Forms, and Geometric Shapes
- Miscellaneous Symbols
- Presentation Forms
- Braille Patterns
- Musical Symbols (Western, Byzantine, & Ancient Greek)



PSI

Efficient Encoding: UTF-16

2007-03-08 Copyright © 2000 - 2007 Paladin Software Incorporated, Inc. 26

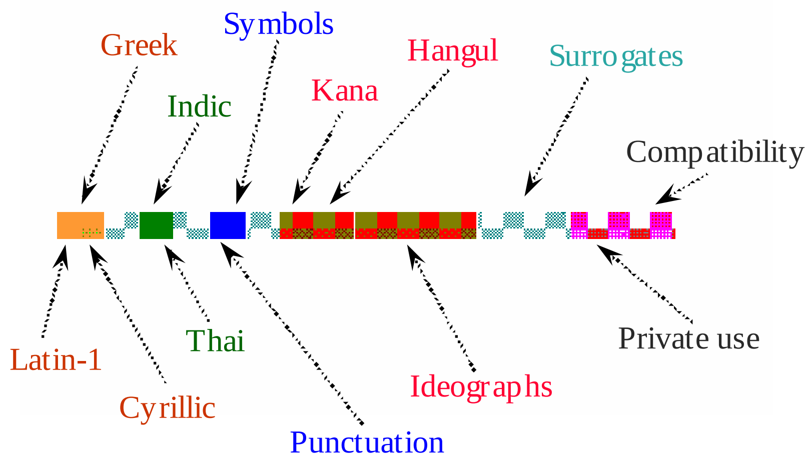


PSI

Count of Unicode Characters

2007-03-08 Copyright © 2000 - 2007 Paladin Software Incorporated, Inc. 27

The Groups — BMP (approximate)

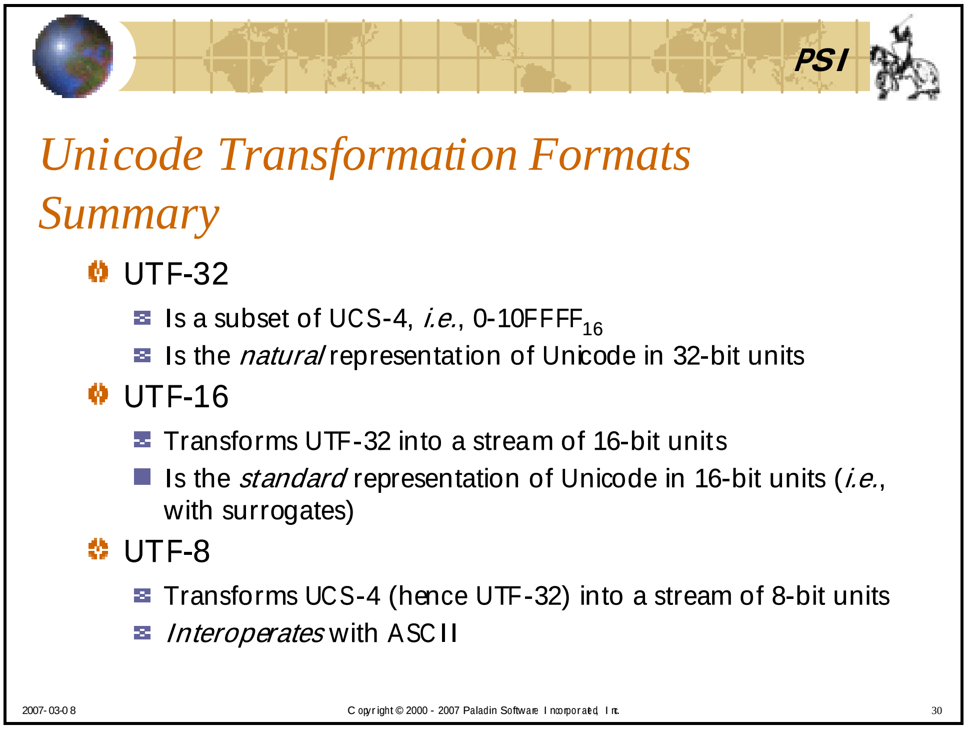


2007-03-08 Copyright © 2000 - 2007 Paladin Software Incorporated, Inc. 28

Interoperable Encoding: UTF-8

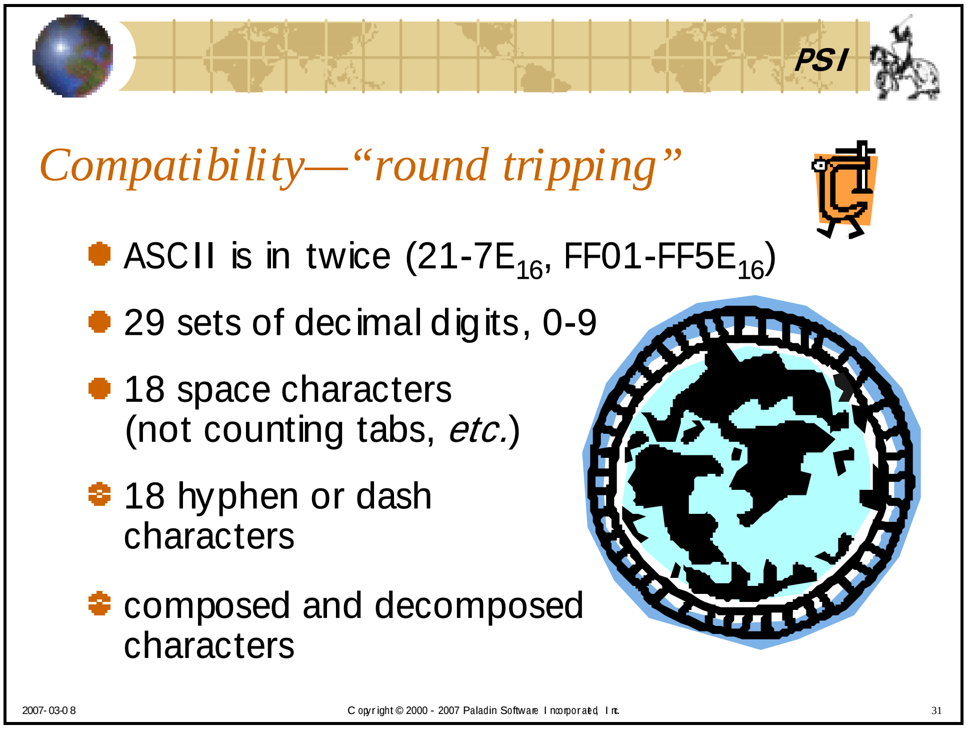
UTF-32	UTF-8	bits
0000007F	0·······	7
000007FF	110····· 10······	11
0000FFFF	1110···· 10······ 10······	16
001FFFFF	11110··· 10······ 10······ 10······	21
03FFFFFF	111110·· 10······ 10······ 10······ 10······	26
7FFFFFFF	1111110· 10······ 10······ 10······ 10······ 10······	31

2007-03-08 Copyright © 2000 - 2007 Paladin Software Incorporated, Inc. 29



Unicode Transformation Formats Summary

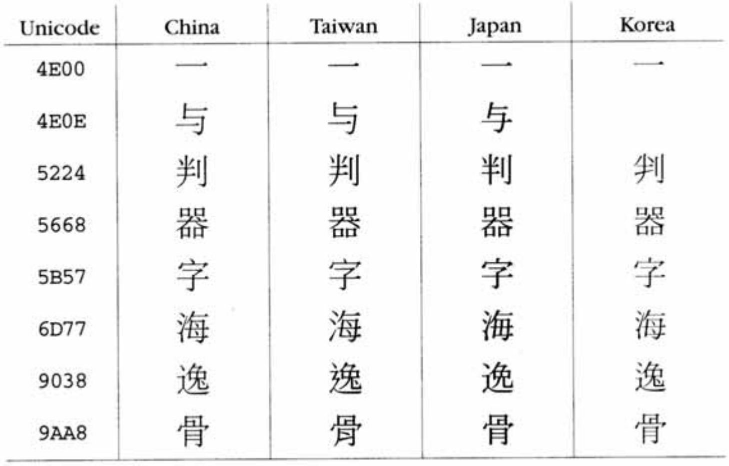
- UTF-32
  - Is a subset of UCS-4, *i.e.*, 0-10FFFF$_{16}$
  - Is the *natural* representation of Unicode in 32-bit units
- UTF-16
  - Transforms UTF-32 into a stream of 16-bit units
  - Is the *standard* representation of Unicode in 16-bit units (*i.e.*, with surrogates)
- UTF-8
  - Transforms UCS-4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



Compatibility—“round tripping”

- ASCII is in twice (21-7E$_{16}$, FF01-FF5E$_{16}$)
- 29 sets of decimal digits, 0-9
- 18 space characters (not counting tabs, *etc.*)
- 18 hyphen or dash characters
- composed and decomposed characters

Han Unification

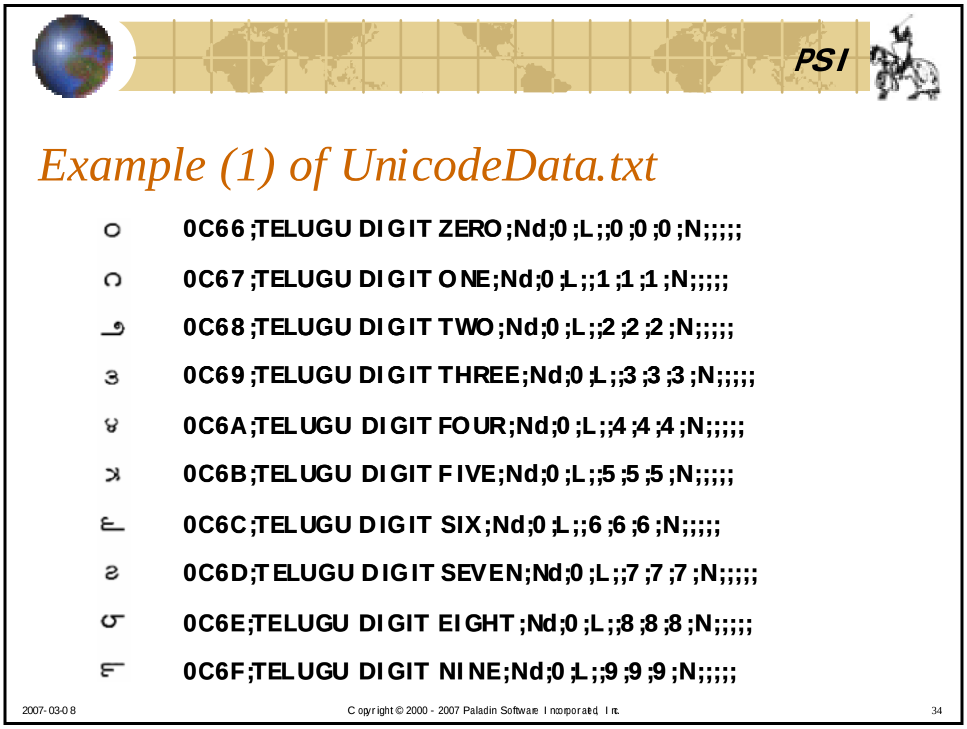


Unicode	China	Taiwan	Japan	Korea
4E00	一	一	一	一
4E0E	与	与	与	
5224	判	判	判	判
5668	器	器	器	器
5B57	字	字	字	字
6D77	海	海	海	海
9038	逸	逸	逸	逸
9AA8	骨	骨	骨	骨

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

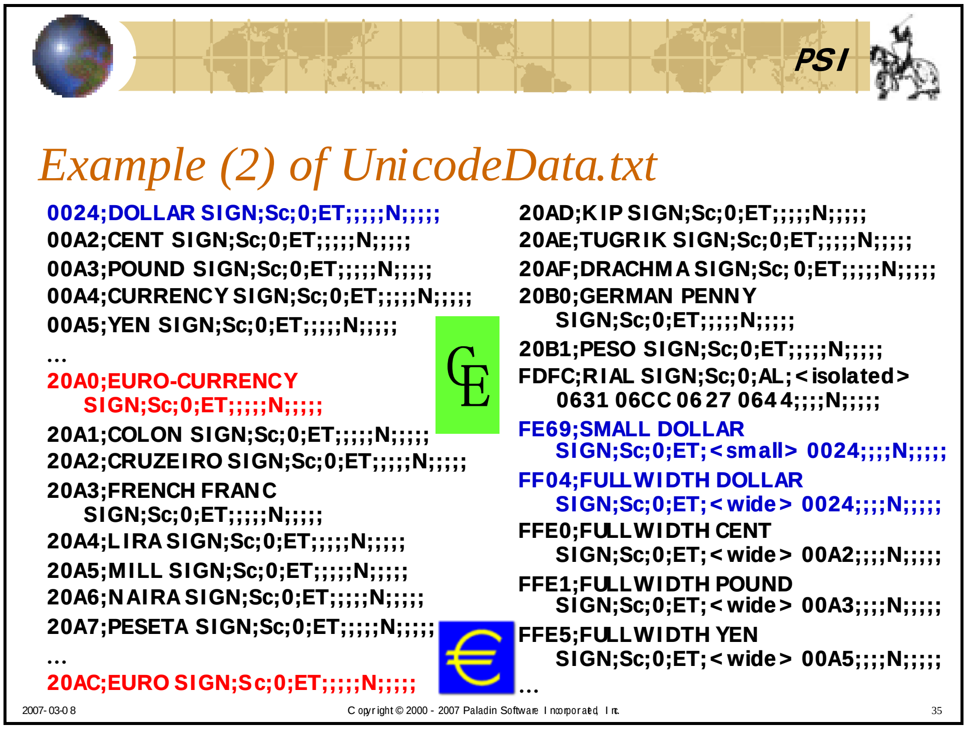
UnicodeData.txt



Example (1) of UnicodeData.txt

```
౦   0C66;TELUGU DIGIT ZERO;Nd;0;L;;0;0;0;N;;;;
౧   0C67;TELUGU DIGIT ONE;Nd;0;L;;1;1;1;N;;;;
౨   0C68;TELUGU DIGIT TWO;Nd;0;L;;2;2;2;N;;;;
౩   0C69;TELUGU DIGIT THREE;Nd;0;L;;3;3;3;N;;;;
౪   0C6A;TELUGU DIGIT FOUR;Nd;0;L;;4;4;4;N;;;;
౫   0C6B;TELUGU DIGIT FIVE;Nd;0;L;;5;5;5;N;;;;
౬   0C6C;TELUGU DIGIT SIX;Nd;0;L;;6;6;6;N;;;;
౭   0C6D;TELUGU DIGIT SEVEN;Nd;0;L;;7;7;7;N;;;;
౮   0C6E;TELUGU DIGIT EIGHT;Nd;0;L;;8;8;8;N;;;;
౯   0C6F;TELUGU DIGIT NINE;Nd;0;L;;9;9;9;N;;;;
```

2007-03-08 Copyright © 2000 - 2007 Paladin Software Incorporated, Inc. 34



Example (2) of UnicodeData.txt

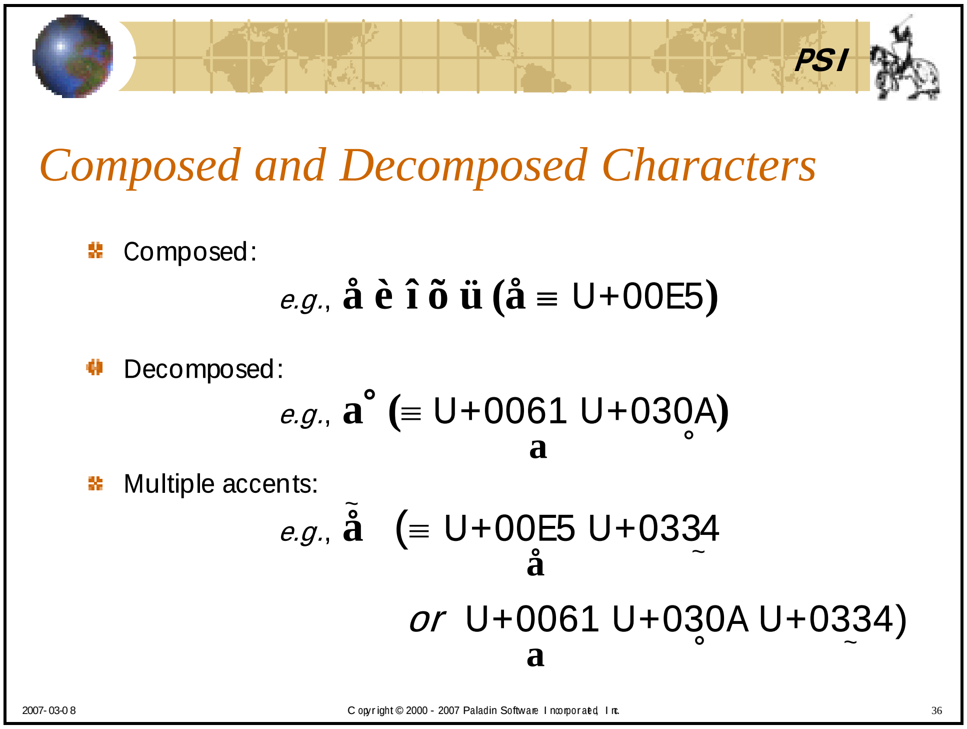
```
0024;DOLLAR SIGN;Sc;0;ET;;;;;N;;;;
00A2;CENT SIGN;Sc;0;ET;;;;;N;;;;
00A3;POUND SIGN;Sc;0;ET;;;;;N;;;;
00A4;CURRENCY SIGN;Sc;0;ET;;;;;N;;;;
00A5;YEN SIGN;Sc;0;ET;;;;;N;;;;
…
20A0;EURO-CURRENCY
    SIGN;Sc;0;ET;;;;;N;;;;
20A1;COLON SIGN;Sc;0;ET;;;;;N;;;;
20A2;CRUZEIRO SIGN;Sc;0;ET;;;;;N;;;;
20A3;FRENCH 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;;;;
…
```

2007-03-08 Copyright © 2000 - 2007 Paladin Software Incorporated, Inc. 35



PSI

Composed and Decomposed Characters

- Composed:

  e.g., **å è î õ ü (å** ≡ U+00E5**)**

- Decomposed:

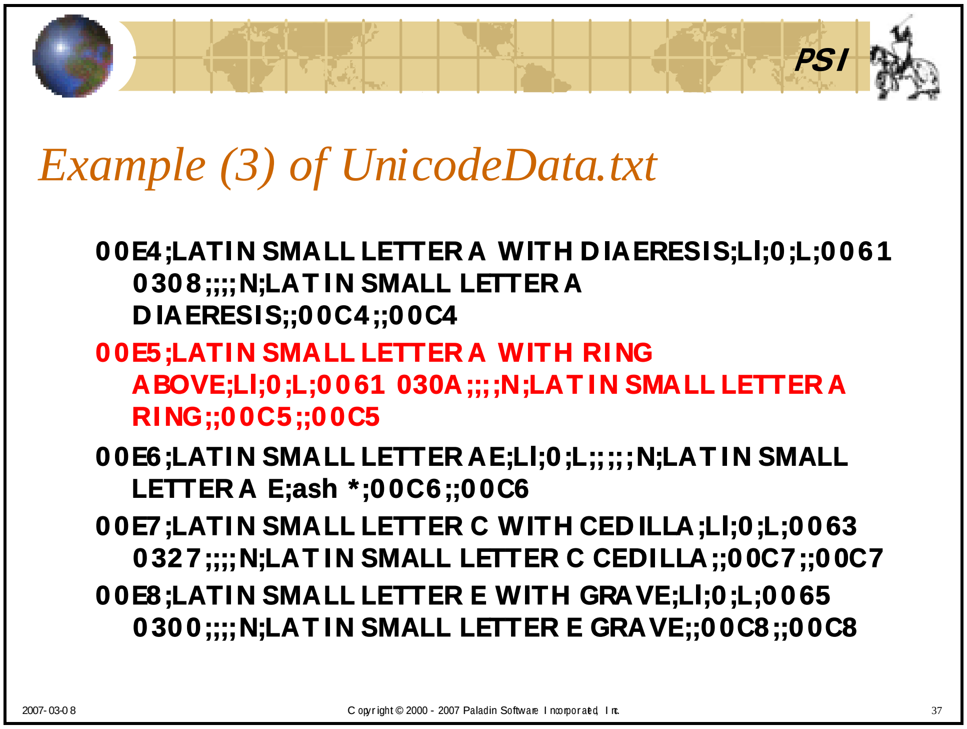
  e.g., **a**° **(**≡ U+0061 U+030A**)**
  a °

- Multiple accents:

  e.g., **å̃** (≡ U+00E5 U+0334
  å ~

  or U+0061 U+030A U+0334)
  a ° ~

2007-03-08 Copyright © 2000 - 2007 Paladin Software Incorporated, Inc. 36



PSI

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

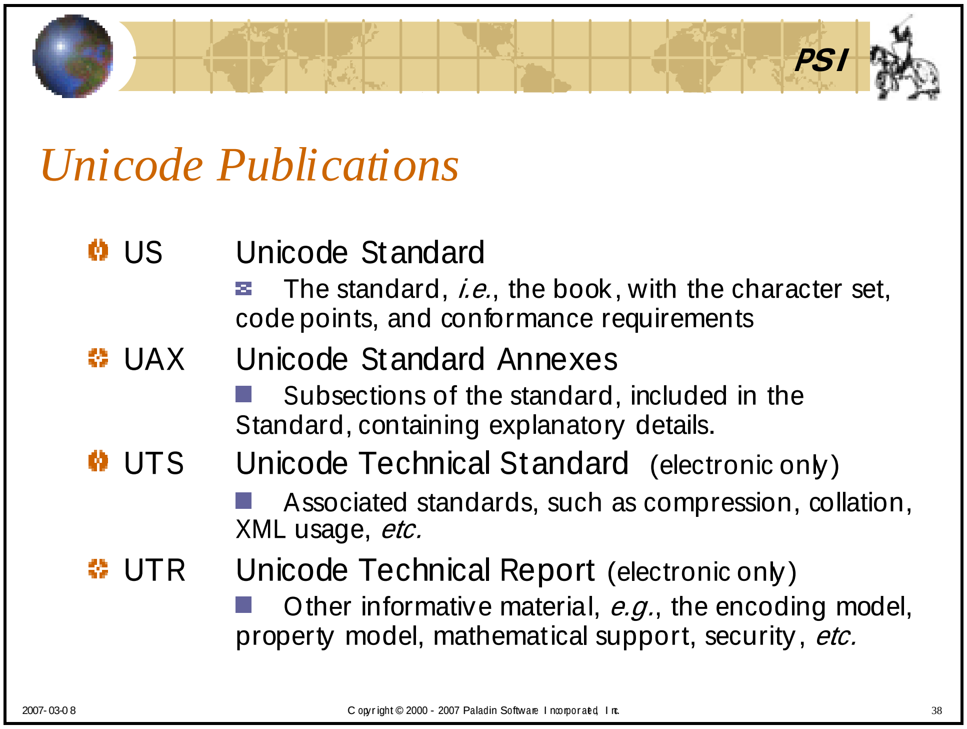
00E5;LATIN SMALL LETTER A WITH RING ABOVE;Ll;0;L;0061 030A;;;;N;LATIN SMALL LETTER A RING;;00C5;;00C5

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

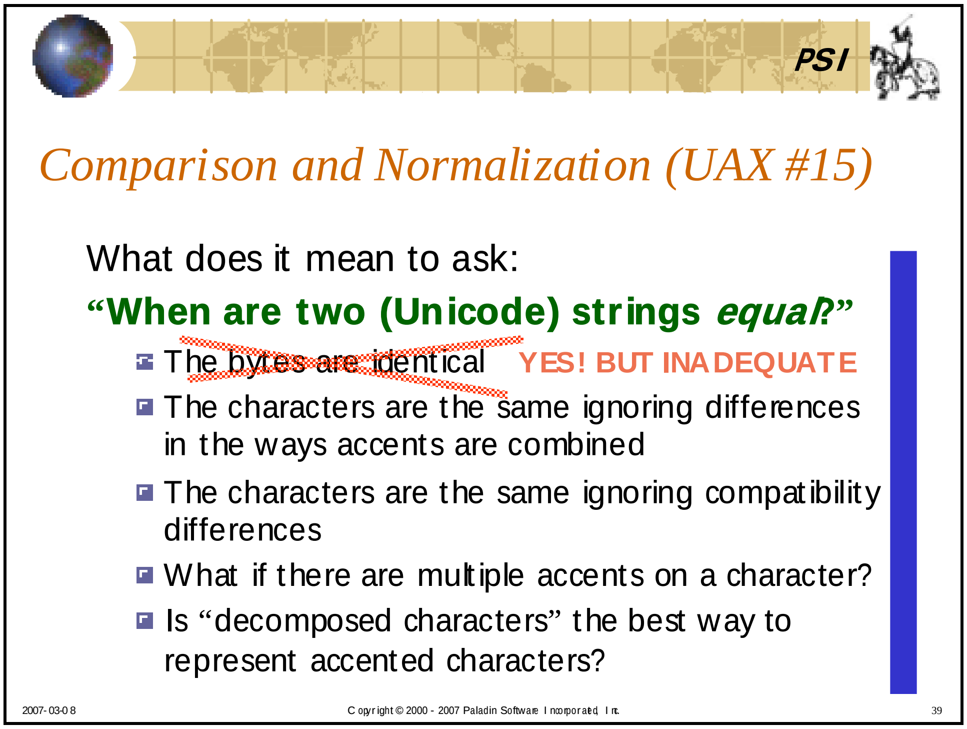
00E8;LATIN SMALL LETTER E WITH GRAVE;Ll;0;L;0065 0300;;;;N;LATIN SMALL LETTER E GRAVE;;00C8;;00C8

2007-03-08 Copyright © 2000 - 2007 Paladin Software Incorporated, Inc. 37



Unicode Publications

- US Unicode Standard
  - The standard, *i.e.*, the book, with the character set, code points, and conformance requirements
- UAX Unicode Standard Annexes
  - Subsections of the standard, included in the Standard, containing explanatory details.
- UTS Unicode Technical Standard (electronic only)
  - Associated standards, such as compression, collation, XML usage, *etc.*
- UTR Unicode Technical Report (electronic only)
  - Other informative material, *e.g.*, the encoding model, property model, mathematical support, security, *etc.*

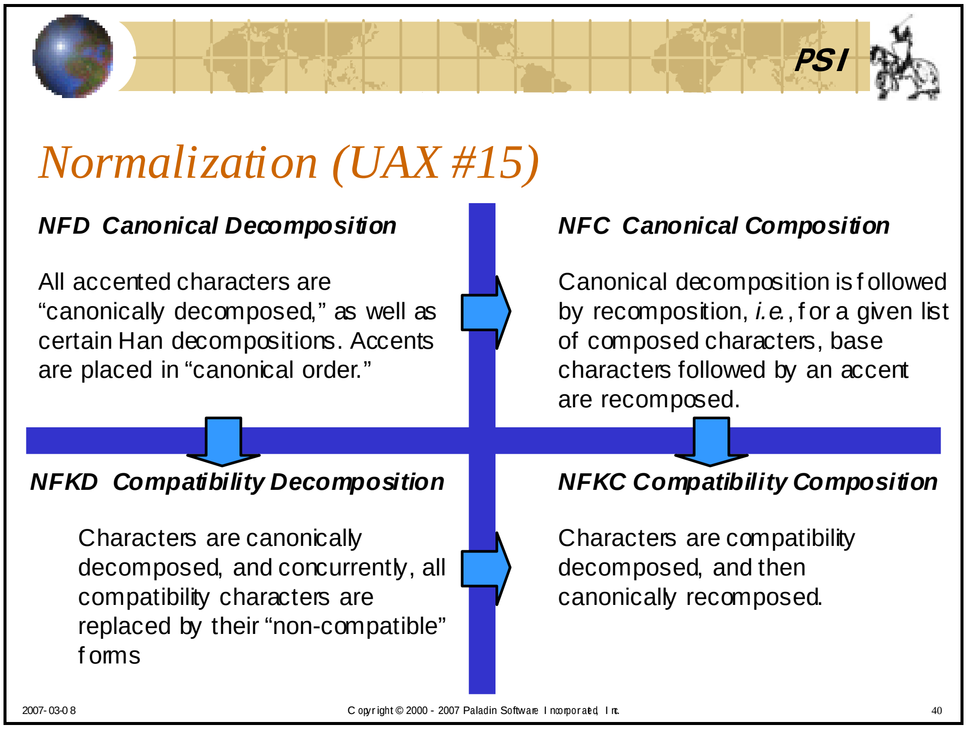


Comparison and Normalization (UAX #15)

What does it mean to ask:

“When are two (Unicode) strings *equal*?”

- ~~The bytes are identical~~ YES! BUT INADEQUATE
- The characters are the same ignoring differences in the ways accents are combined
- The characters are the same ignoring compatibility differences
- What if there are multiple accents on a character?
- Is “decomposed characters” the best way to represent accented characters?



PSI

Normalization (UAX #15)

NFD Canonical Decomposition

All accented characters are "canonically decomposed," as well as certain Han decompositions. Accents are placed in "canonical order."

NFC Canonical Composition

Canonical decomposition is followed by recomposition, *i.e.*, for a given list of composed characters, base characters followed by an accent are recomposed.

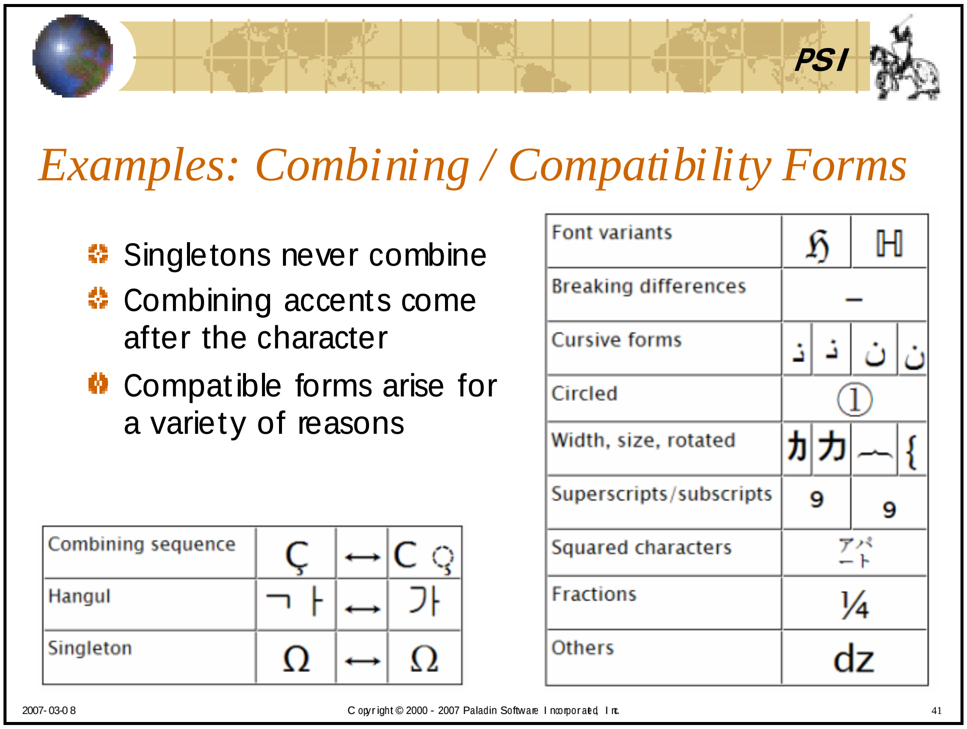
NFKD Compatibility Decomposition

Characters are canonically decomposed, and concurrently, all compatibility characters are replaced by their "non-compatible" forms

NFKC Compatibility Composition

Characters are compatibility decomposed, and then canonically recomposed.

2007-03-08 Copyright © 2000 - 2007 Paladin Software Incorporated, Inc. 40



PSI

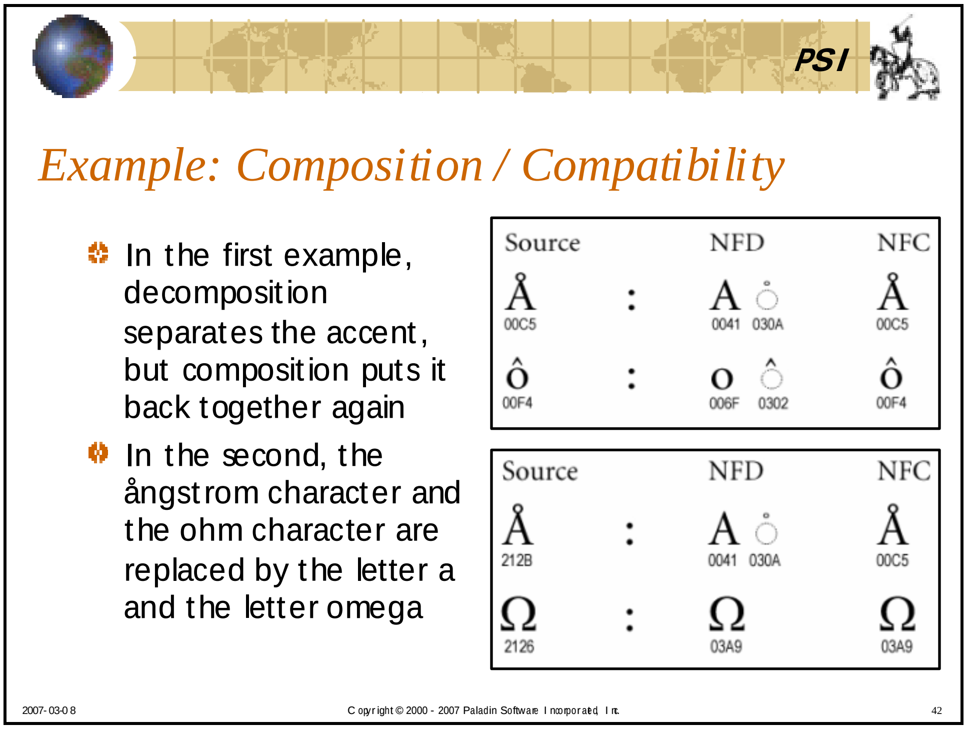
Examples: Combining / Compatibility Forms

- Singletons never combine
- Combining accents come after the character
- Compatible forms arise for a variety of reasons

Combining sequence	Ç	↔	C ̧
Hangul	ㄱ ㅏ	↔	가
Singleton	Ω	↔	Ω

Font variants	ℌ		ℍ	
Breaking differences	‑			
Cursive forms	ﻨ	ﻧ	ﻥ	ﻦ
Circled	①			
Width, size, rotated	ｶ	カ	︵	{
Superscripts/subscripts	⁹		₉	
Squared characters	㌂			
Fractions	¼			
Others	ǳ			

2007-03-08 Copyright © 2000 - 2007 Paladin Software Incorporated, Inc. 41



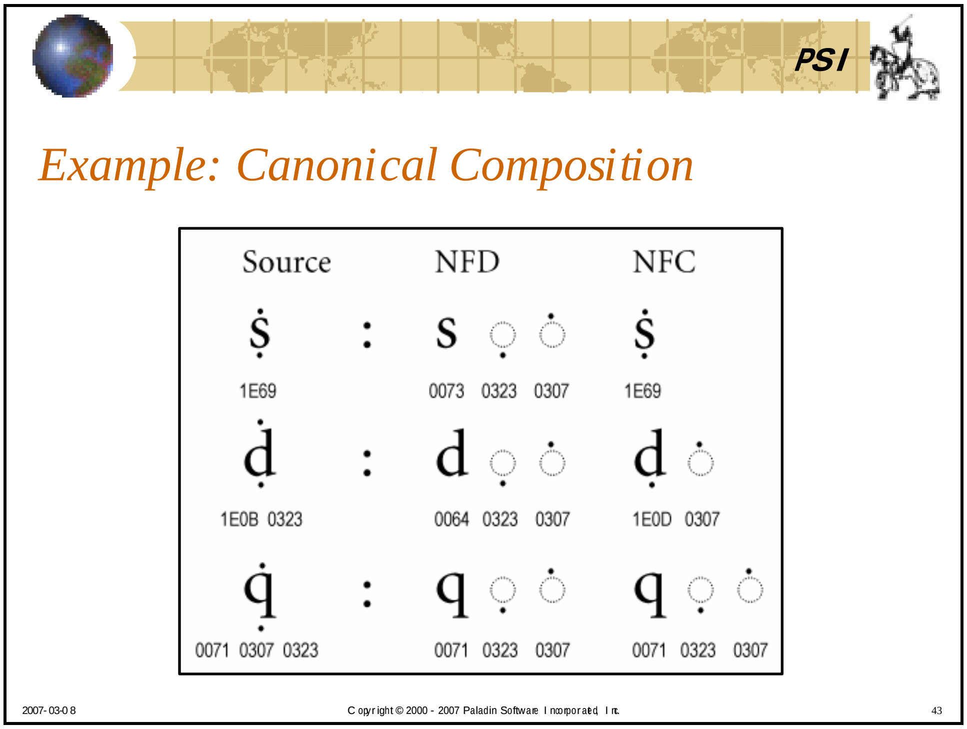
Example: Composition / Compatibility

- In the first example, decomposition separates the accent, but composition puts it back together again
- In the second, the ångstrom character and the ohm character are replaced by the letter a and the letter omega

Source		NFD	NFC
Å 00C5	:	A ◌̊ 0041 030A	Å 00C5
ô 00F4	:	o ◌̂ 006F 0302	ô 00F4

Source		NFD	NFC
Å 212B	:	A ◌̊ 0041 030A	Å 00C5
Ω 2126	:	Ω 03A9	Ω 03A9

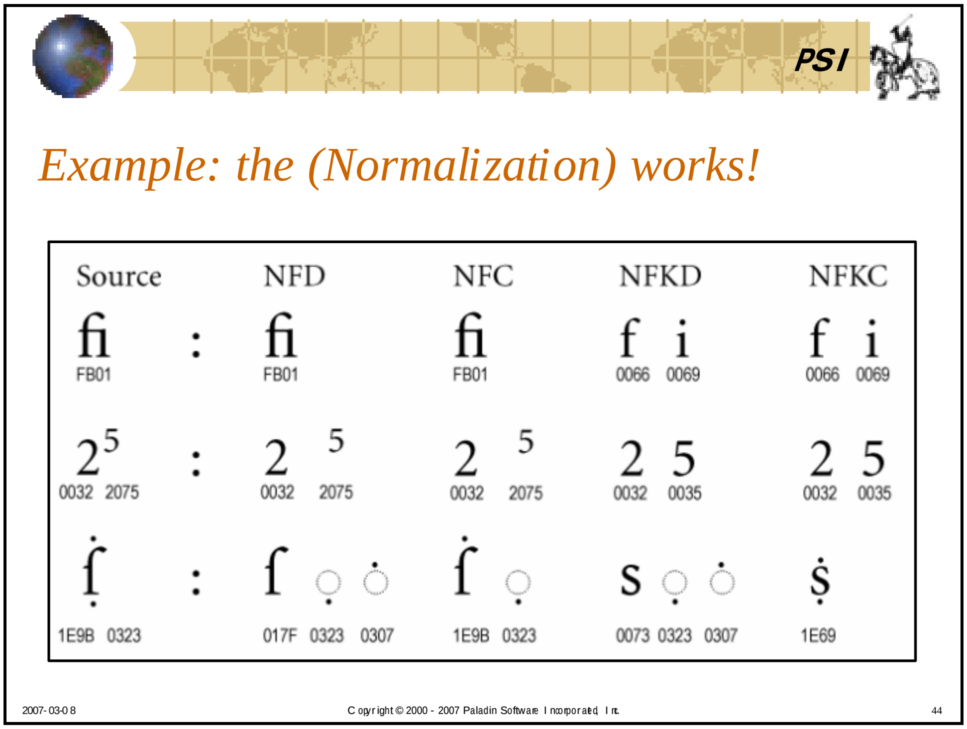
2007-03-08 Copyright © 2000 - 2007 Paladin Software Incorporated, Inc. 42



Example: Canonical Composition

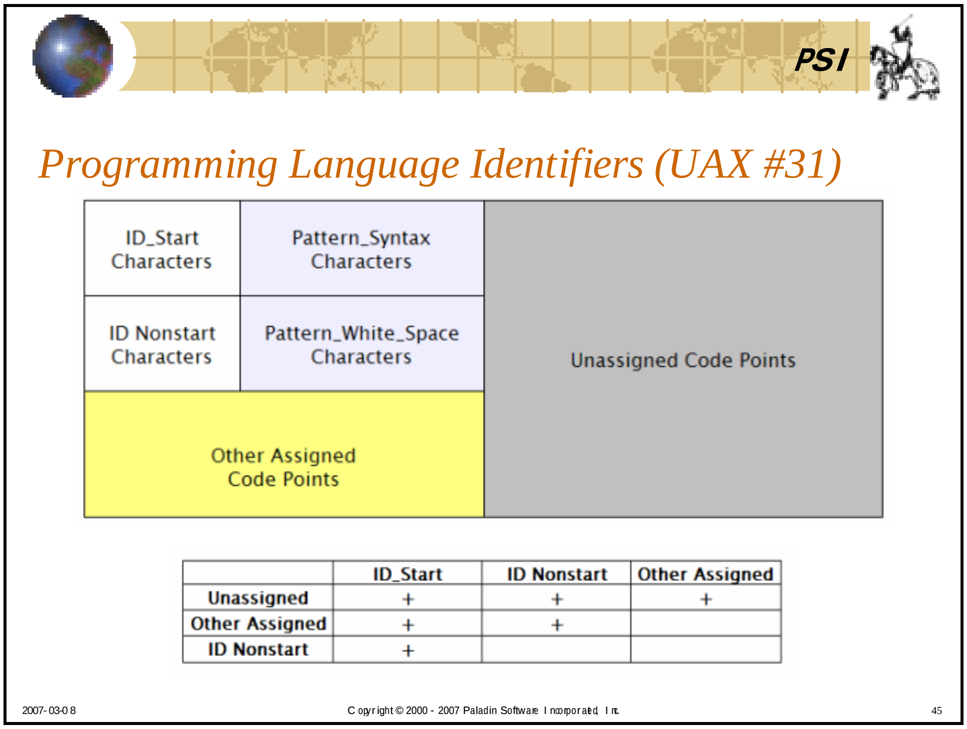
Source		NFD	NFC
ṩ 1E69	:	s ◌̣ ◌̇ 0073 0323 0307	ṩ 1E69
ḍ̇ 1E0B 0323	:	d ◌̣ ◌̇ 0064 0323 0307	ḍ ◌̇ 1E0D 0307
q̣̇ 0071 0307 0323	:	q ◌̣ ◌̇ 0071 0323 0307	q ◌̣ ◌̇ 0071 0323 0307

2007-03-08 Copyright © 2000 - 2007 Paladin Software Incorporated, Inc. 43



Example: the (Normalization) works!

Source		NFD	NFC	NFKD	NFKC
ﬁ FB01	:	ﬁ FB01	ﬁ FB01	f i 0066 0069	f i 0066 0069
2^{5} 0032 2075	:	2 ^{5} 0032 2075	2 ^{5} 0032 2075	2 5 0032 0035	2 5 0032 0035
ẛ̣ 1E9B 0323	:	ſ ◌̣ ◌̇ 017F 0323 0307	ẛ ◌̣ 1E9B 0323	s ◌̣ ◌̇ 0073 0323 0307	ṩ 1E69



Programming Language Identifiers (UAX #31)

ID_Start Characters

Pattern_Syntax Characters

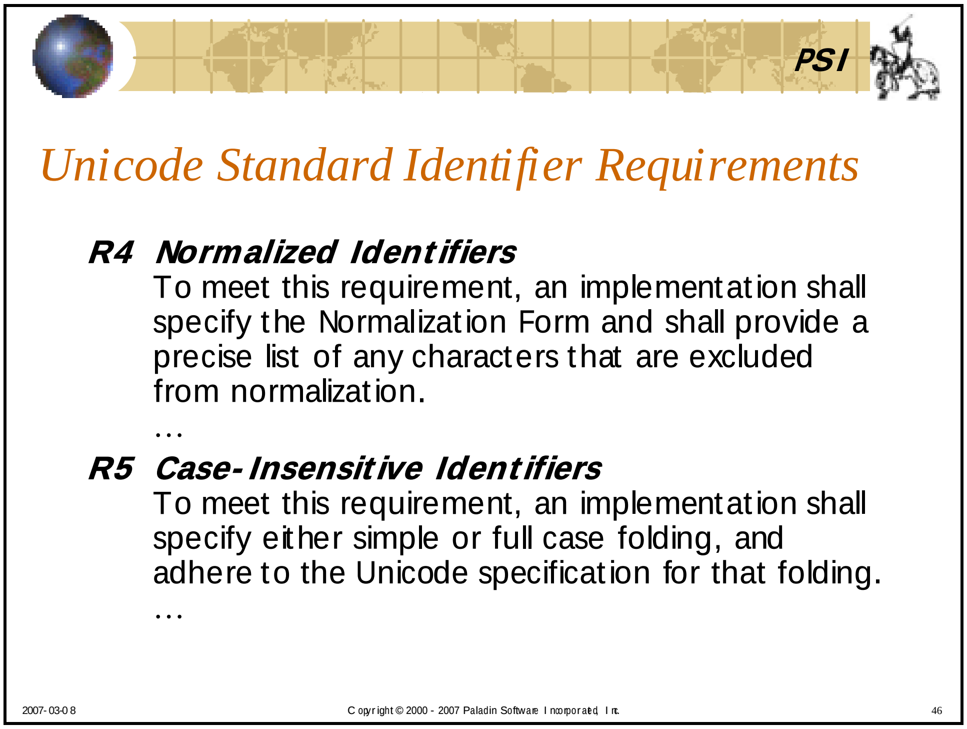
ID Nonstart Characters

Pattern_White_Space Characters

Unassigned Code Points

Other Assigned Code Points

	ID_Start	ID Nonstart	Other Assigned
Unassigned	+	+	+
Other Assigned	+	+	
ID Nonstart	+		



PSI

Unicode Standard Identifier Requirements

R4 Normalized Identifiers

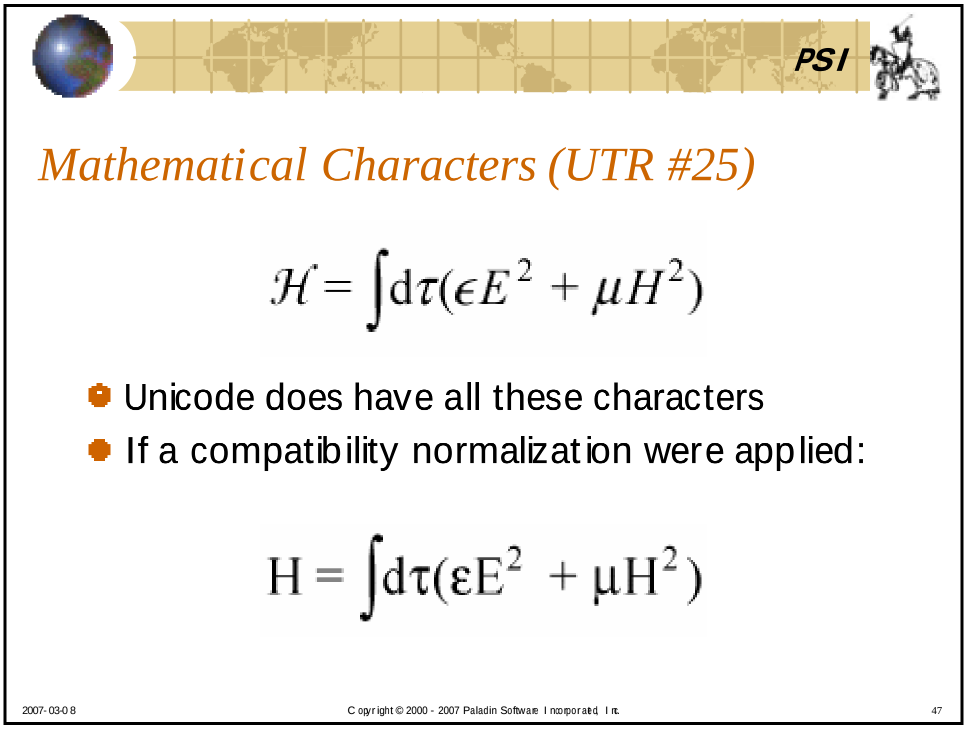
To meet this requirement, an implementation shall specify the Normalization Form and shall provide a precise list of any characters that are excluded from normalization.

…

R5 Case-Insensitive Identifiers

To meet this requirement, an implementation shall specify either simple or full case folding, and adhere to the Unicode specification for that folding.

…



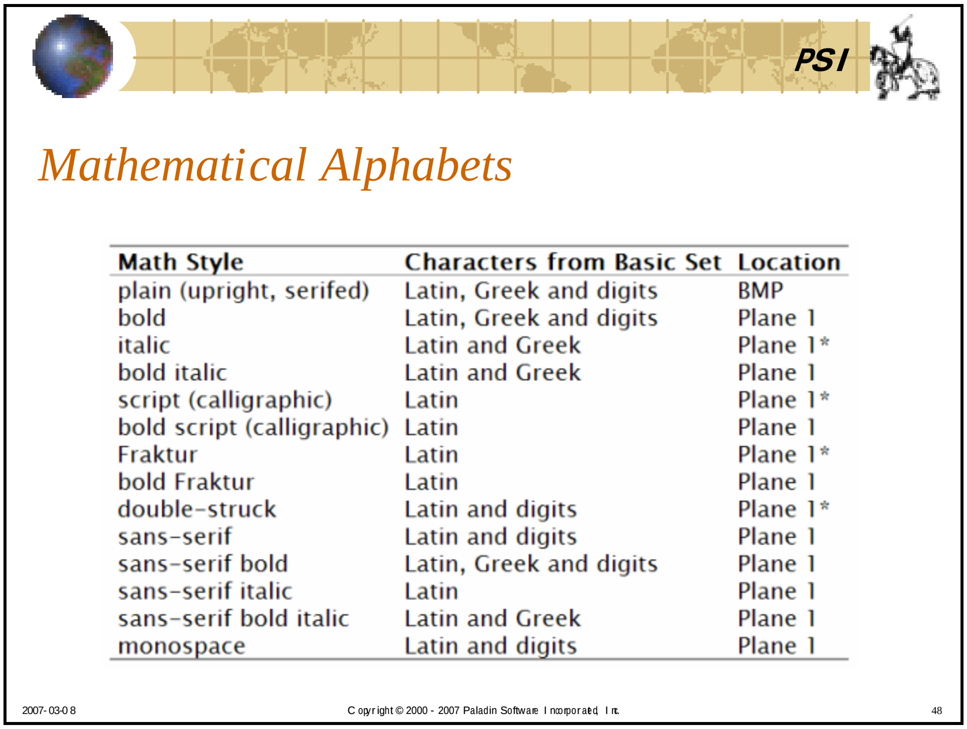
PSI

Mathematical Characters (UTR #25)

$$\mathcal{H} = \int d\tau(\epsilon E^2 + \mu H^2)$$

- Unicode does have all these characters
- If a compatibility normalization were applied:

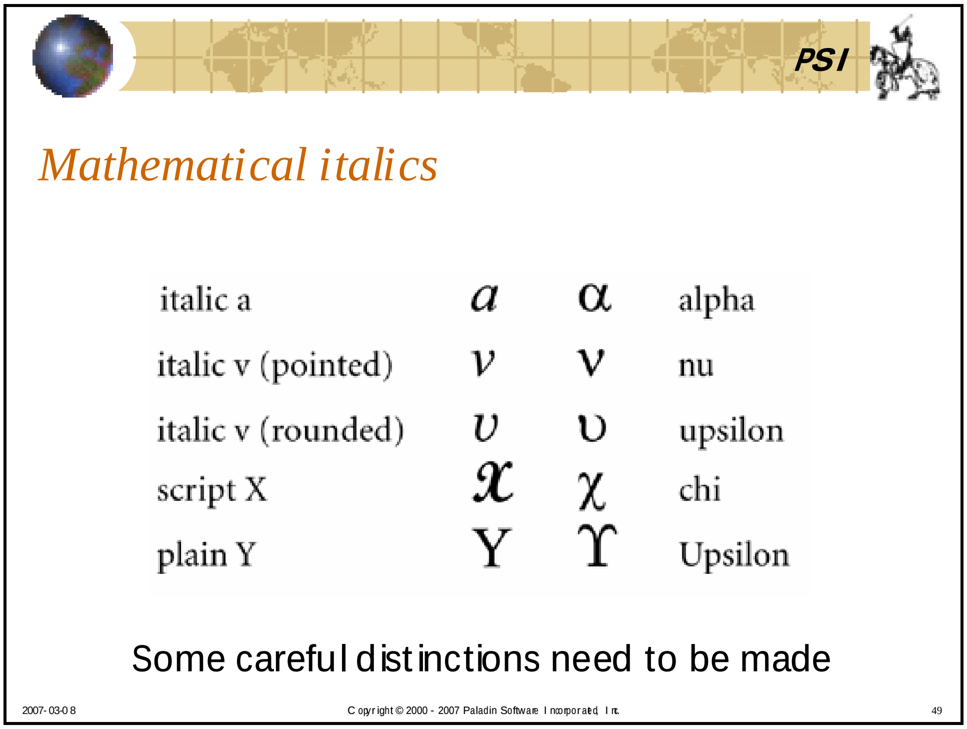
$$\mathrm{H} = \int \mathrm{d}\tau(\varepsilon \mathrm{E}^2 + \mu \mathrm{H}^2)$$



PSI

Mathematical Alphabets

Math Style	Characters from Basic Set	Location
plain (upright, serifed)	Latin, Greek and digits	BMP
bold	Latin, Greek and digits	Plane 1
italic	Latin and Greek	Plane 1*
bold italic	Latin and Greek	Plane 1
script (calligraphic)	Latin	Plane 1*
bold script (calligraphic)	Latin	Plane 1
Fraktur	Latin	Plane 1*
bold Fraktur	Latin	Plane 1
double-struck	Latin and digits	Plane 1*
sans-serif	Latin and digits	Plane 1
sans-serif bold	Latin, Greek and digits	Plane 1
sans-serif italic	Latin	Plane 1
sans-serif bold italic	Latin and Greek	Plane 1
monospace	Latin and digits	Plane 1

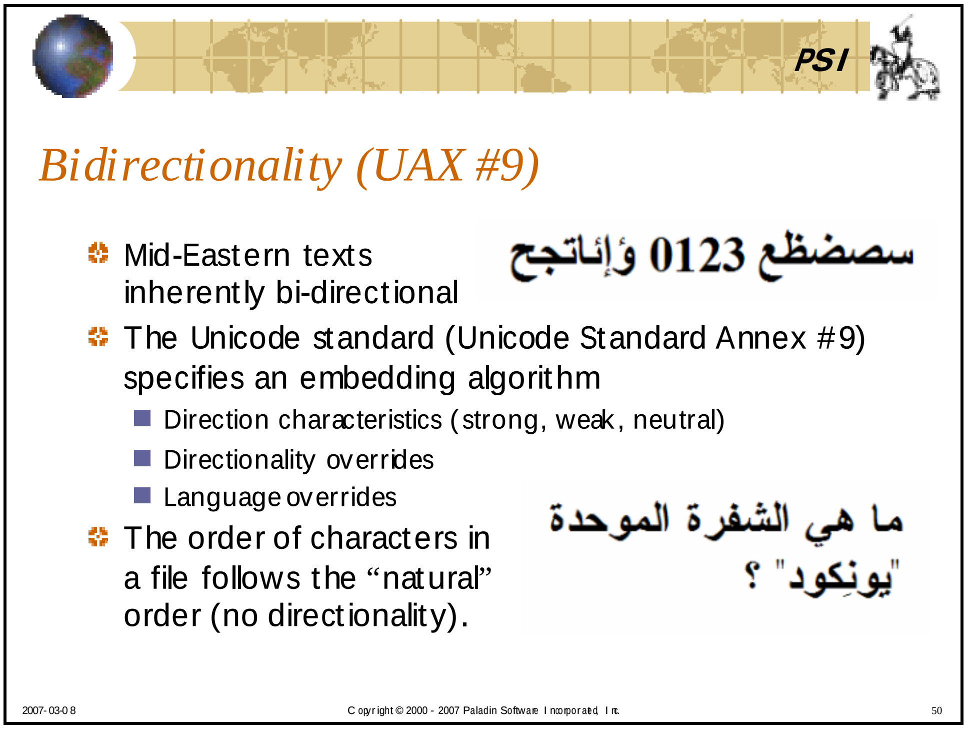


PSI

Mathematical italics

italic a	𝑎	α	alpha
italic v (pointed)	𝑣	ν	nu
italic v (rounded)	𝜐	υ	upsilon
script X	𝒳	χ	chi
plain Y	Y	ϒ	Upsilon

Some careful distinctions need to be made



PSI

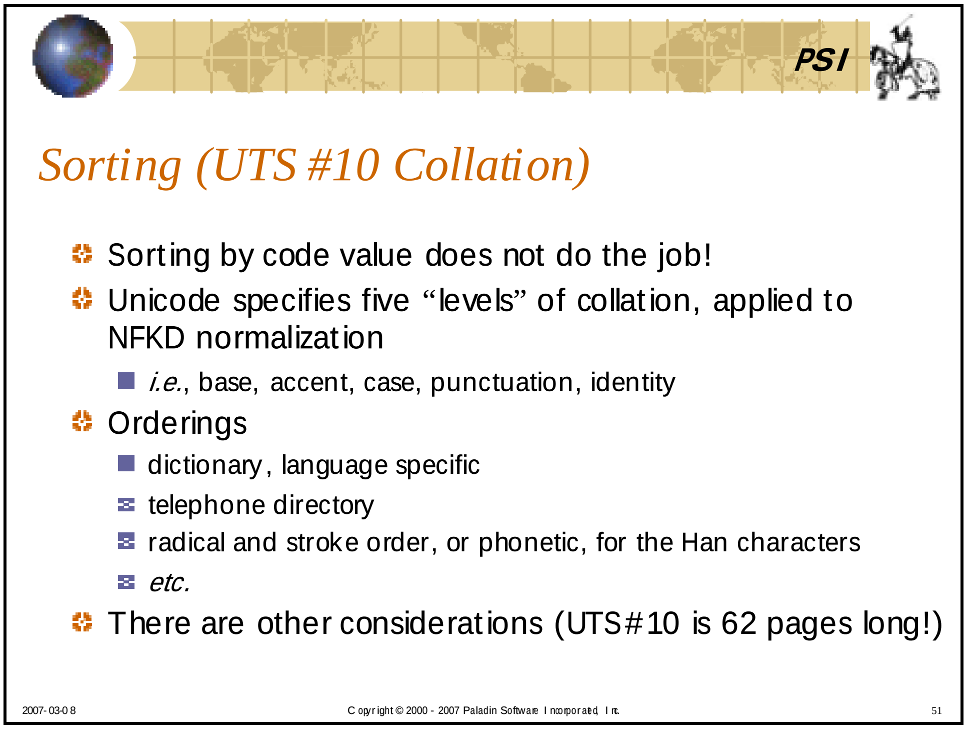
Bidirectionality (UAX #9)

- Mid-Eastern texts inherently bi-directional

سصضظع 0123 وؤإئاتجح

- 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).

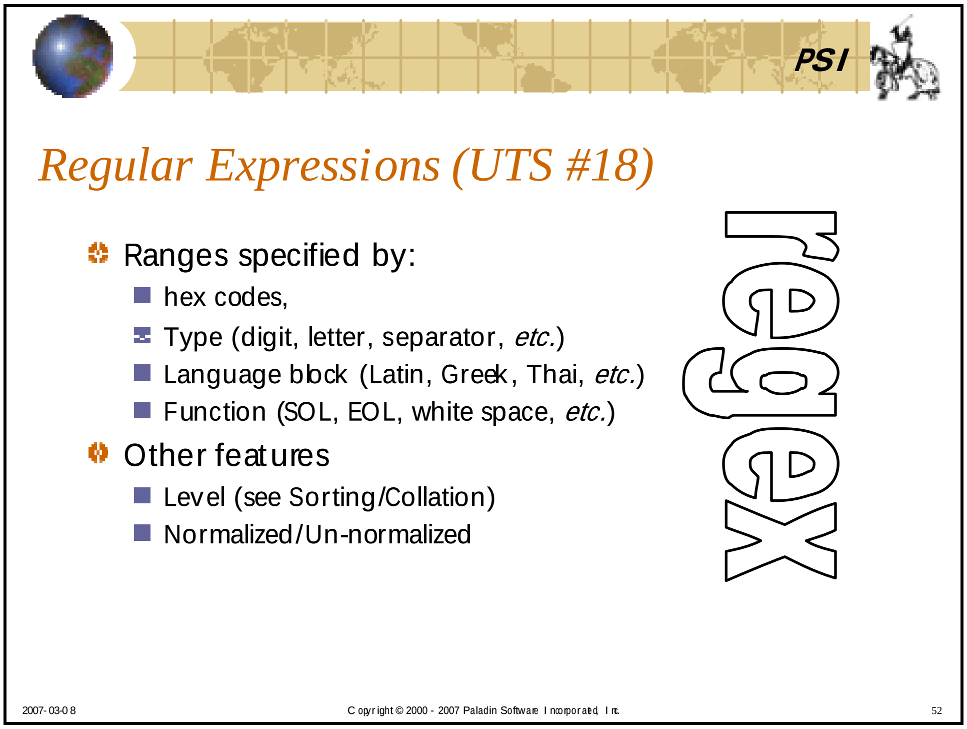
ما هي الشفرة الموحدة
"يونِكود" ؟



PSI

Sorting (UTS #10 Collation)

- Sorting by code value does not do the job!
- Unicode specifies five "levels" of collation, applied to NFKD normalization
  - *i.e.*, base, accent, case, punctuation, identity
- Orderings
  - dictionary, language specific
  - telephone directory
  - radical and stroke order, or phonetic, for the Han characters
  - *etc.*
- There are other considerations (UTS#10 is 62 pages long!)

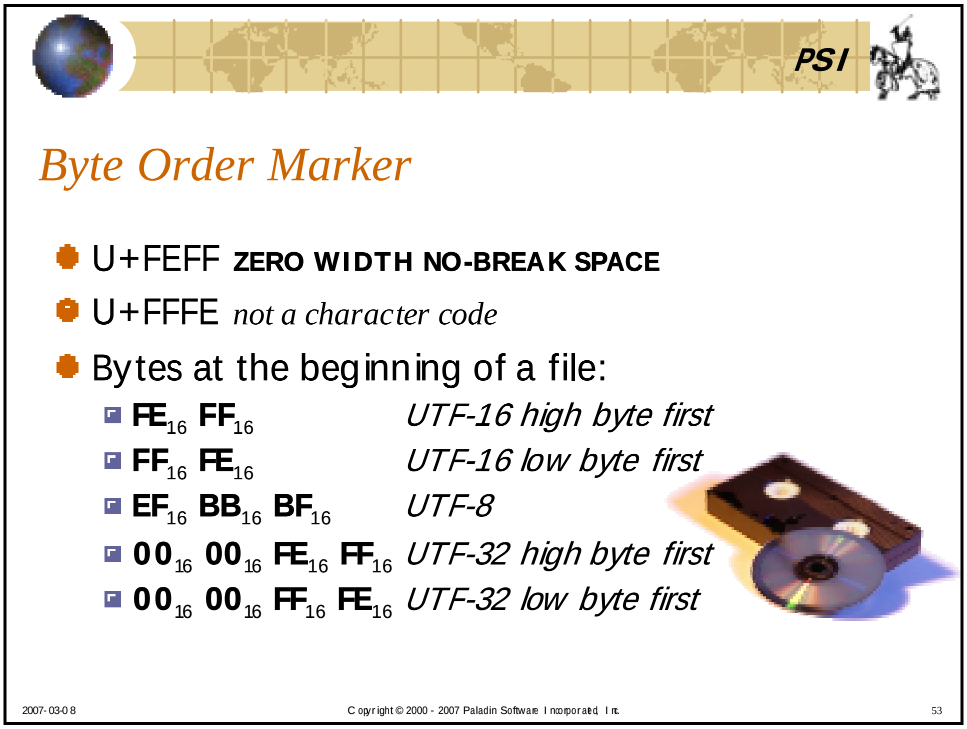


PSI

Regular Expressions (UTS #18)

regex

- 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/Collation)
  - Normalized/Un-normalized



PSI

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$_{16}$ **FE**$_{16}$	*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**$_{16}$ **FE**$_{16}$	*UTF-32 low byte first*

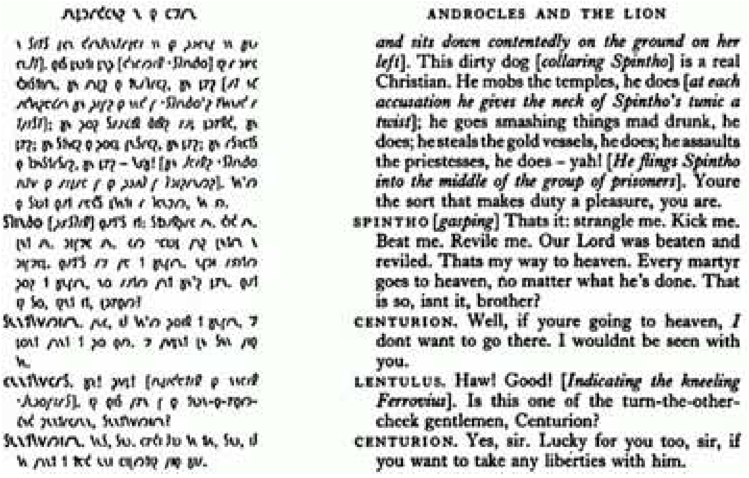
New Scripts in Version 4.0

- 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
- Total 1226 new

4.0 Statistics

Graphic	96,245
Format	137
Control	65
Private Use	137,468
Noncharacter	66

Shavian Script



ANDROCLES AND THE LION

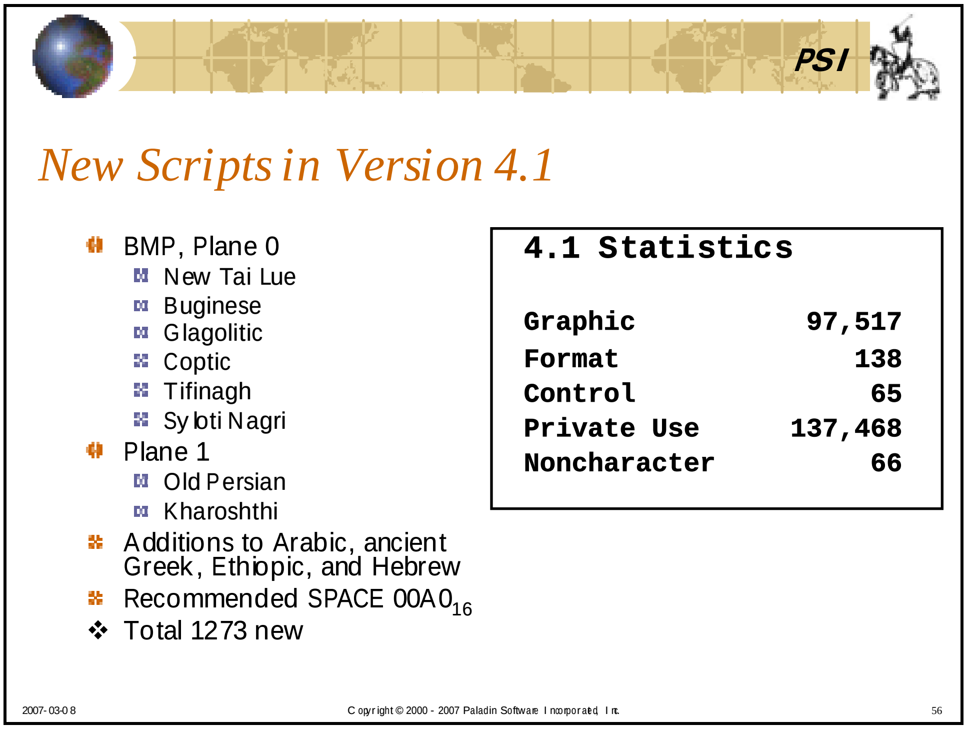
and sits down contentedly on the ground on her left]. This dirty dog [*collaring Spintho*] is a real Christian. He mobs the temples, he does [*at each accusation he gives the neck of Spintho's tunic a twist*]; he goes smashing things mad drunk, he does; he steals the gold vessels, he does; he assaults 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.

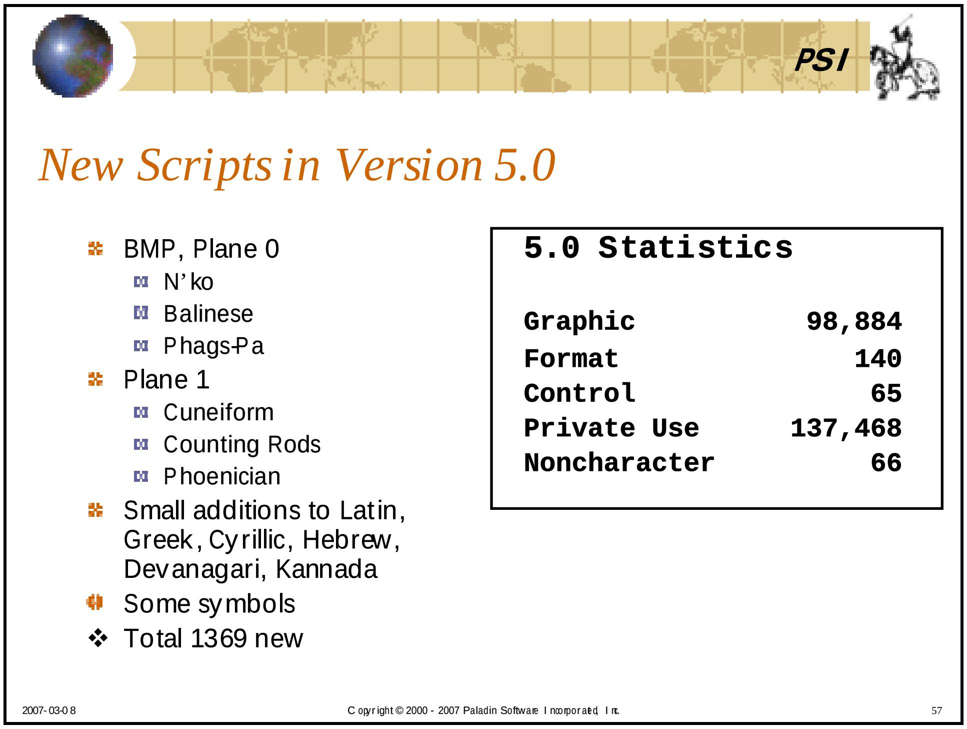


PSI

New Scripts in Version 4.1

- BMP, Plane 0
  - New Tai Lue
  - Buginese
  - Glagolitic
  - Coptic
  - Tifinagh
  - Syloti Nagri
- Plane 1
  - Old Persian
  - Kharoshthi
- Additions to Arabic, ancient Greek, Ethiopic, and Hebrew
- Recommended SPACE 00A0$_{16}$
- Total 1273 new

4.1 Statistics	
Graphic	97,517
Format	138
Control	65
Private Use	137,468
Noncharacter	66

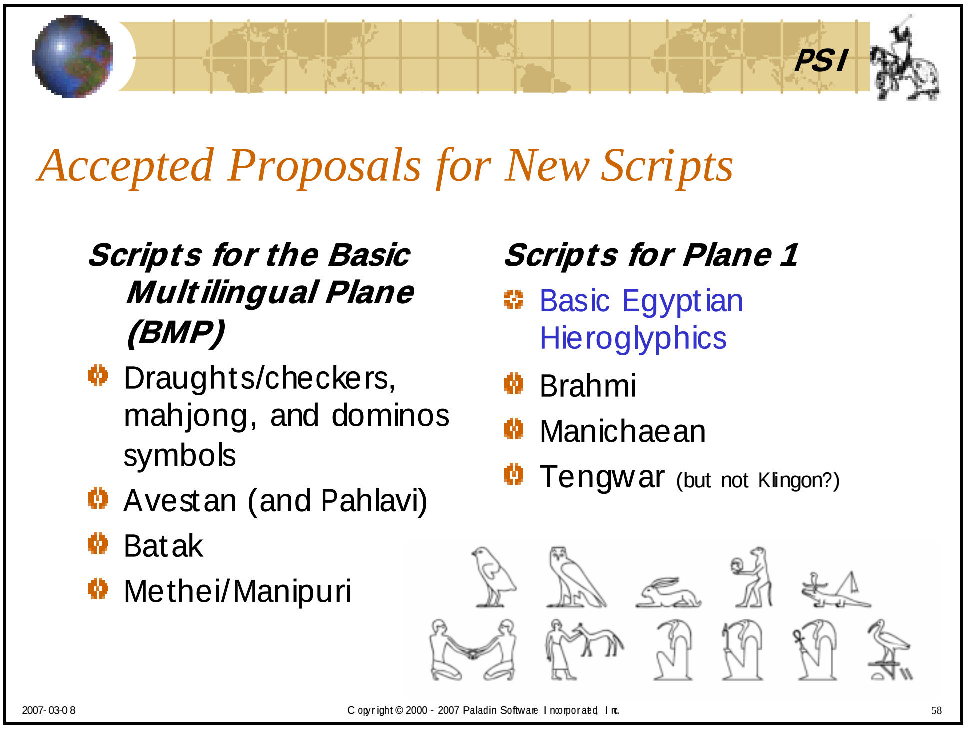


PSI

New Scripts in Version 5.0

- BMP, Plane 0
  - N'ko
  - Balinese
  - Phags-Pa
- Plane 1
  - Cuneiform
  - Counting Rods
  - Phoenician
- Small additions to Latin, Greek, Cyrillic, Hebrew, Devanagari, Kannada
- Some symbols
- Total 1369 new

5.0 Statistics	
Graphic	98,884
Format	140
Control	65
Private Use	137,468
Noncharacter	66



PSI

Accepted Proposals for New Scripts

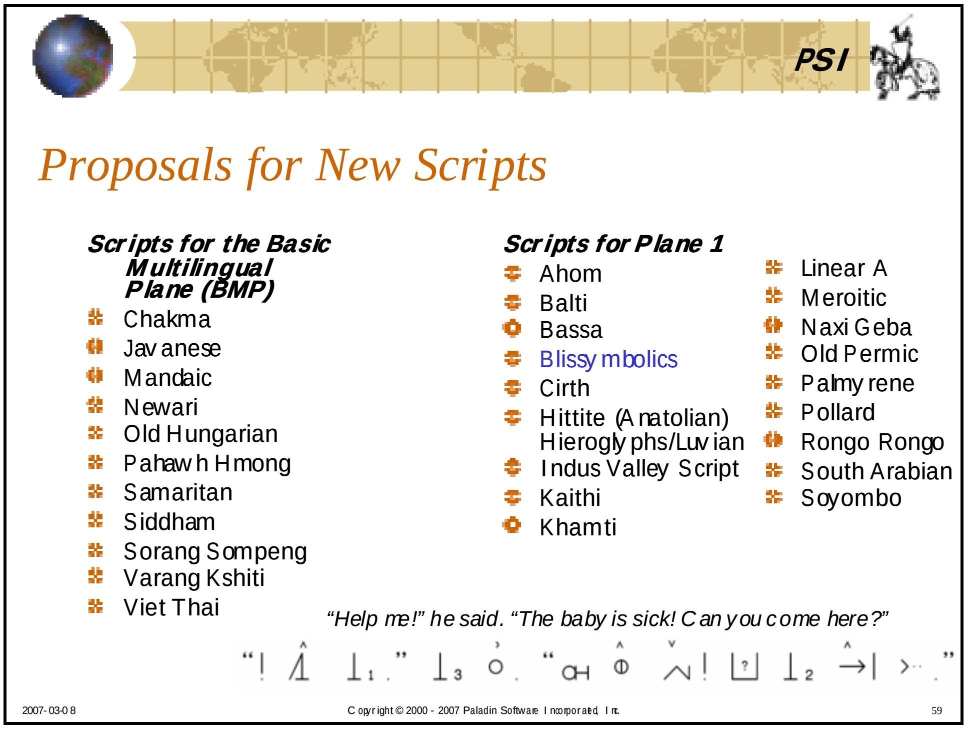
Scripts for the Basic Multilingual Plane (BMP)

- Draughts/checkers, mahjong, and dominos symbols
- Avestan (and Pahlavi)
- Batak
- Methei/Manipuri

Scripts for Plane 1

- Basic Egyptian Hieroglyphics
- Brahmi
- Manichaean
- Tengwar (but not Klingon?)

2007-03-08 Copyright © 2000 - 2007 Paladin Software Incorporated, Inc. 58



PSI

Proposals for New Scripts

Scripts for the Basic Multilingual Plane (BMP)

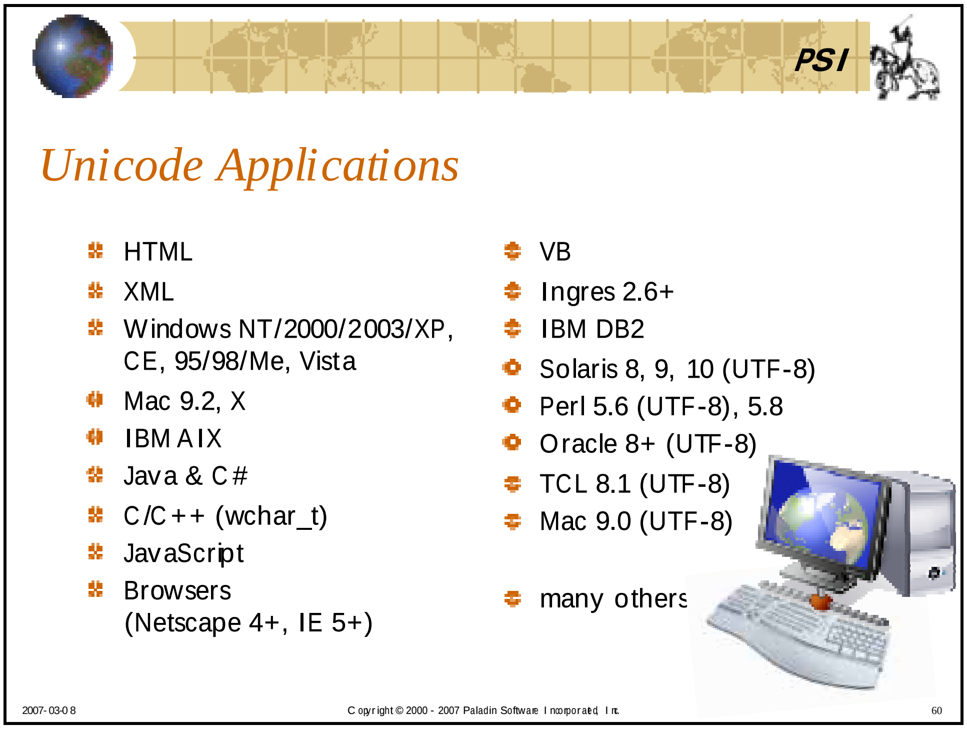
- Chakma
- Javanese
- Mandaic
- Newari
- Old Hungarian
- Pahawh Hmong
- Samaritan
- Siddham
- Sorang Sompeng
- Varang Kshiti
- Viet Thai

Scripts for Plane 1

- Ahom
- Balti
- Bassa
- Blissymbolics
- Cirth
- Hittite (Anatolian) Hieroglyphs/Luvian
- Indus Valley Script
- Kaithi
- Khamti
- Linear A
- Meroitic
- Naxi Geba
- Old Permic
- Palmyrene
- Pollard
- Rongo Rongo
- South Arabian
- Soyombo

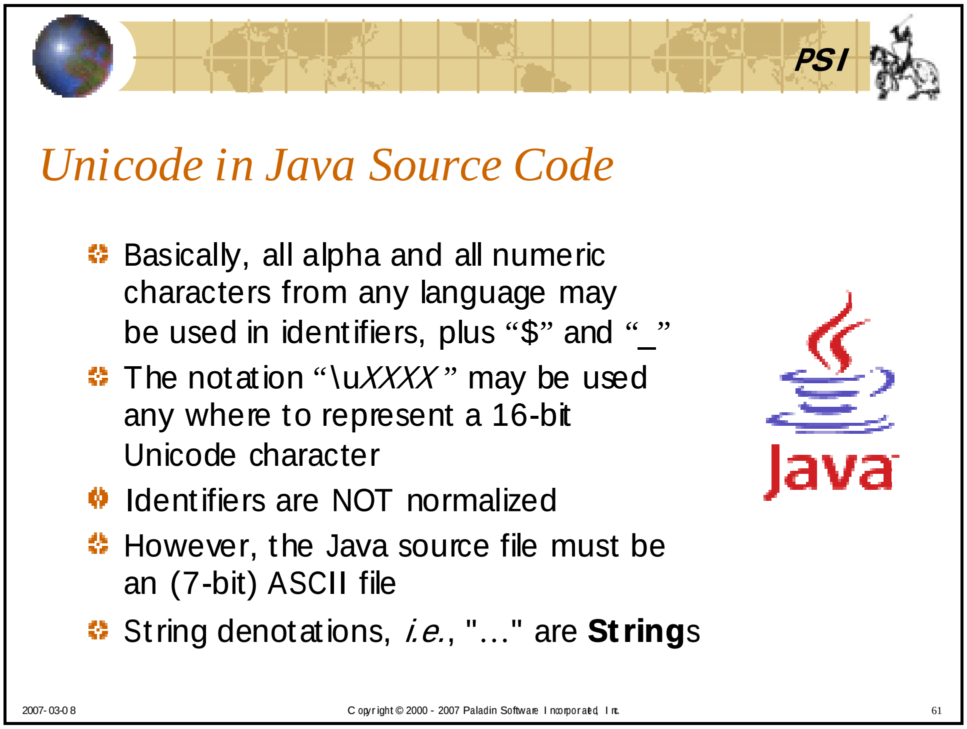
"Help me!" he said. "The baby is sick! Can you come here?"

2007-03-08 Copyright © 2000 - 2007 Paladin Software Incorporated, Inc. 59



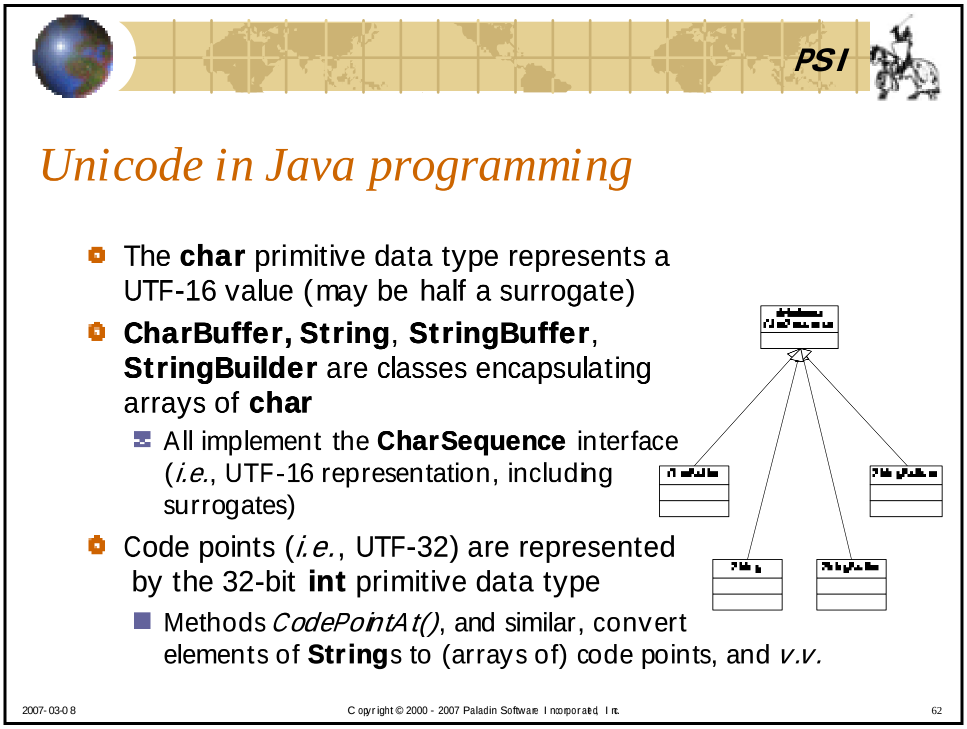
Unicode Applications

- HTML
- XML
- Windows NT/2000/2003/XP, CE, 95/98/Me, Vista
- Mac 9.2, X
- IBM AIX
- Java & C#
- C/C++ (wchar_t)
- JavaScript
- Browsers (Netscape 4+, IE 5+)
- VB
- Ingres 2.6+
- IBM DB2
- Solaris 8, 9, 10 (UTF-8)
- Perl 5.6 (UTF-8), 5.8
- Oracle 8+ (UTF-8)
- TCL 8.1 (UTF-8)
- Mac 9.0 (UTF-8)
- many others



Unicode in Java Source Code

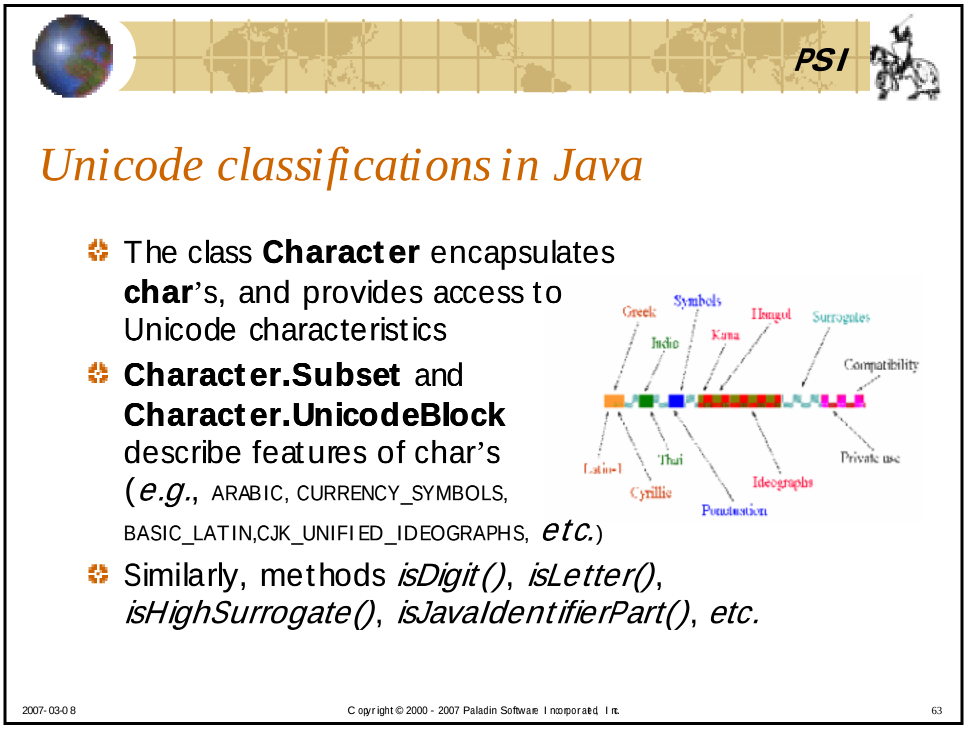
- Basically, all alpha and all numeric characters from any language may be used in identifiers, plus “$” and “_”
- The notation “\u*XXXX*” may be used any where to represent a 16-bit Unicode character
- Identifiers are NOT normalized
- However, the Java source file must be an (7-bit) ASCII file
- String denotations, *i.e.*, "..." are **Strings**



PSI

Unicode in Java programming

- The **char** primitive data type represents a UTF-16 value (may be half a surrogate)
- **CharBuffer, String**, **StringBuffer**, **StringBuilder** are classes encapsulating arrays of **char**
  - All implement the **CharSequence** interface (*i.e.*, UTF-16 representation, including surrogates)
- Code points (*i.e.*, UTF-32) are represented by the 32-bit **int** primitive data type
  - Methods *CodePointAt()*, and similar, convert elements of **Strings** to (arrays of) code points, and *v.v.*

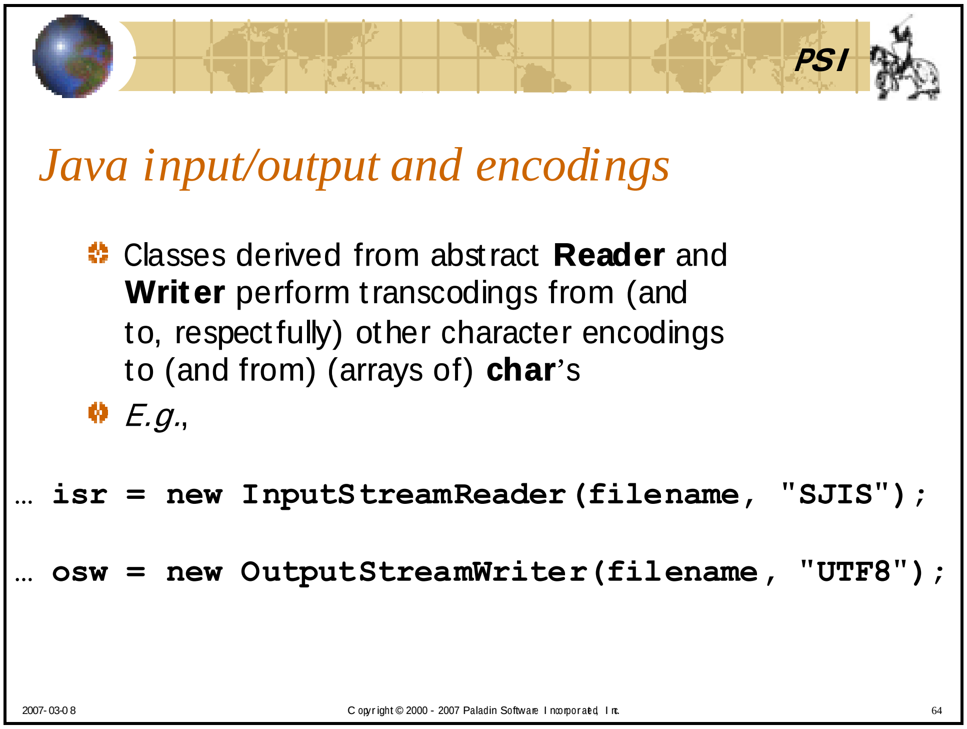


PSI

Unicode classifications in Java

- The class **Character** encapsulates **char**'s, and provides access to Unicode characteristics
- **Character.Subset** and **Character.UnicodeBlock** describe features of char's (*e.g.*, ARABIC, CURRENCY_SYMBOLS, BASIC_LATIN,CJK_UNIFIED_IDEOGRAPHS, *etc.*)
- Similarly, methods *isDigit()*, *isLetter()*, *isHighSurrogate()*, *isJavaIdentifierPart()*, *etc.*

Greek
Symbols
Hangul
Surrogates
Indic
Kana
Compatibility
Latin-1
Thai
Cyrillic
Punctuation
Ideographs
Private use



PSI

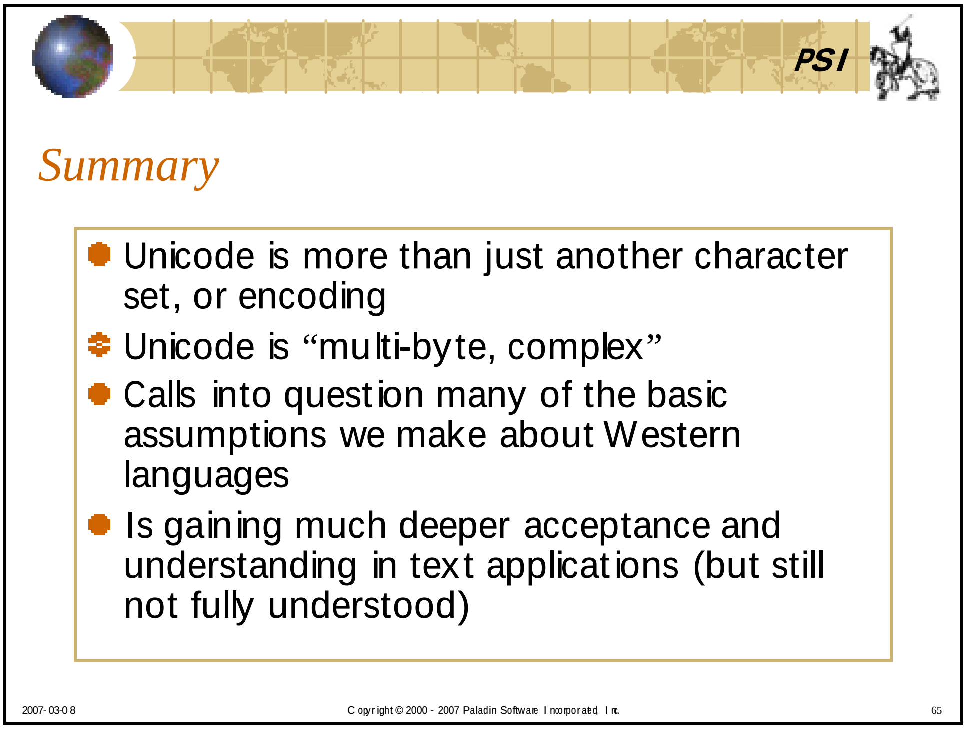
Java input/output and encodings

- Classes derived from abstract **Reader** and **Writer** perform transcodings from (and to, respectfully) other character encodings to (and from) (arrays of) **char**’s
- *E.g.*,

```
… isr = new InputStreamReader(filename, "SJIS");

… osw = new OutputStreamWriter(filename, "UTF8");
```

2007-03-08 Copyright © 2000 - 2007 Paladin Software Incorporated, Inc. 64

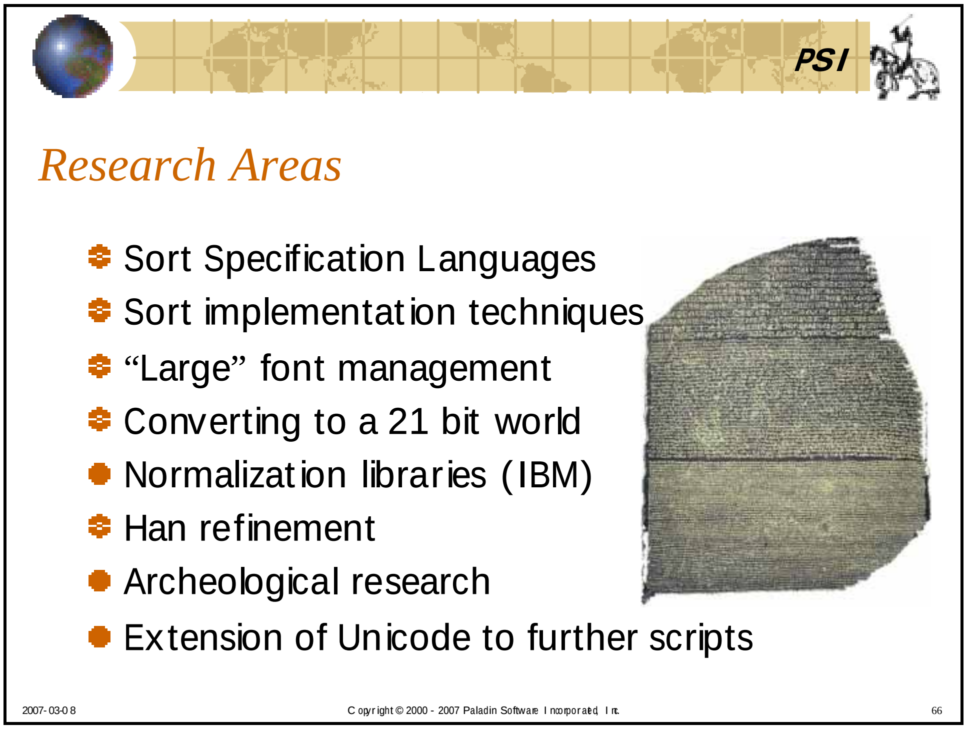


PSI

Summary

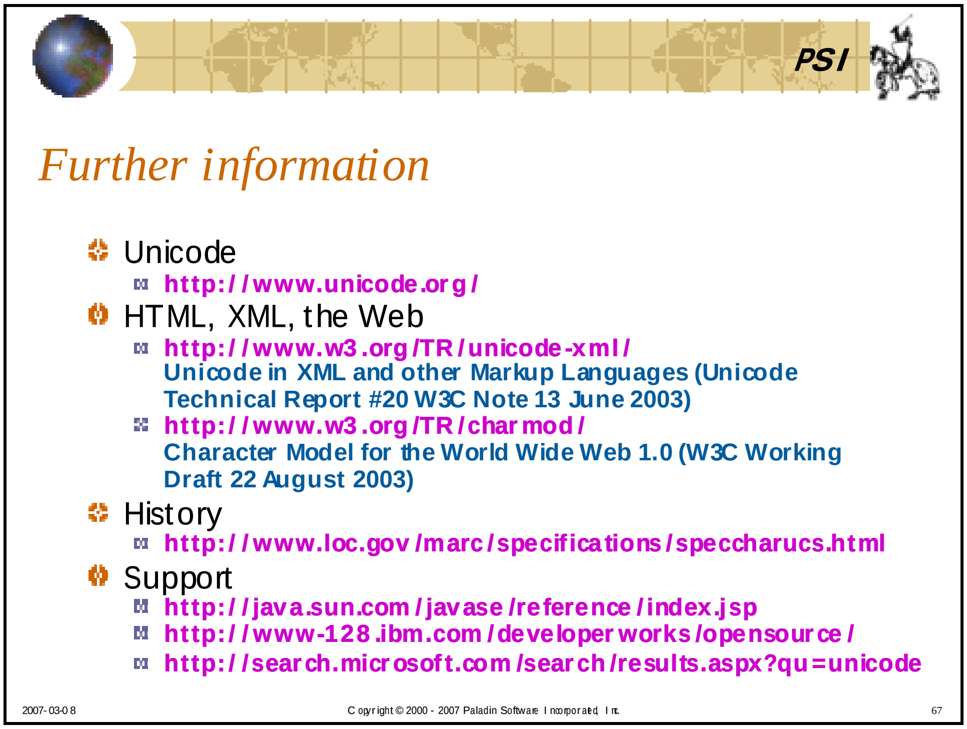
- Unicode is more than just another character set, or encoding
- Unicode is “multi-byte, complex”
- Calls into question many of the basic assumptions we make about Western languages
- Is gaining much deeper acceptance and understanding in text applications (but still not fully understood)

2007-03-08 Copyright © 2000 - 2007 Paladin Software Incorporated, Inc. 65



Research Areas

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



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/javase/reference/index.jsp
  - http://www-128.ibm.com/developerworks/opensource/
  - http://search.microsoft.com/search/results.aspx?qu=unicode