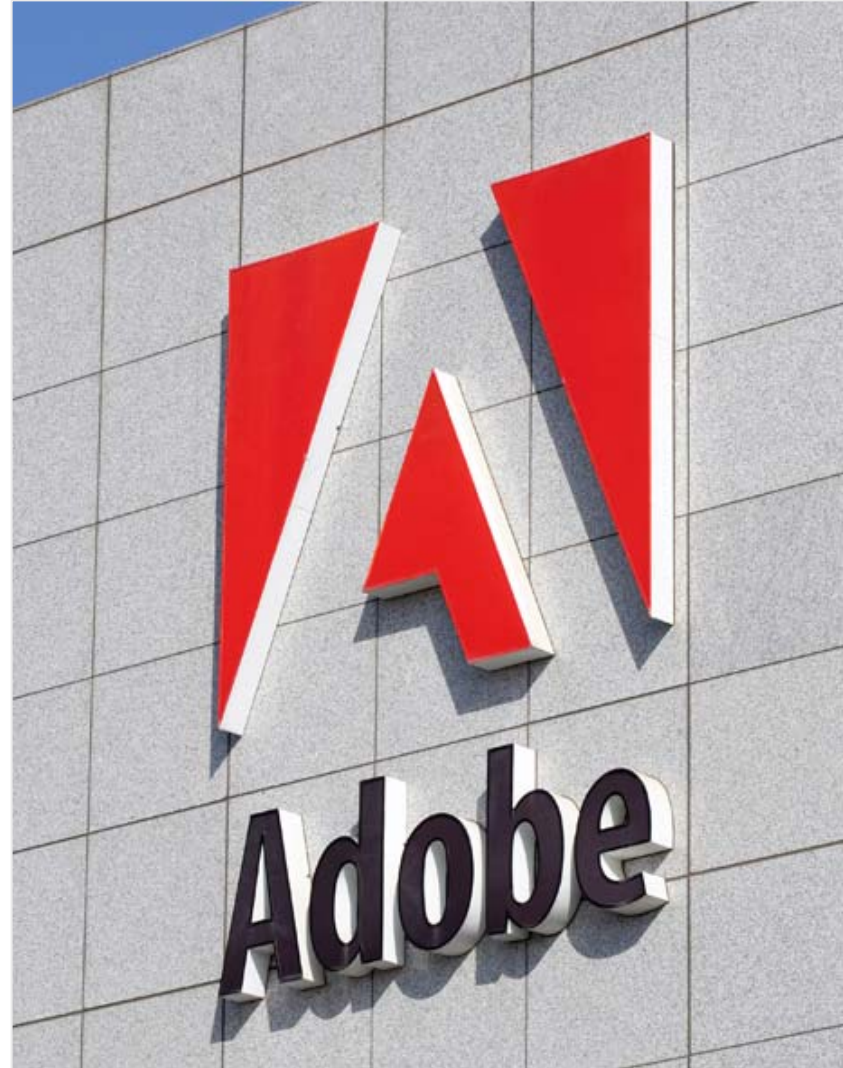


Character Set Journeys

Thomas Phinney

Program Manager
Fonts & SING Technologies

29 September 2006



Agenda

- Definitions
- Encodings before computers
- Computer-based encodings
- Dynamically extending fonts

Definitions

- **Encoding**
 - Digital representation of the basic elements of language (letters, ideographs, etc.)
 - Notation can be binary, decimal, hexadecimal, etc.
- **Codepoint**
 - The number or code for a given character
- **Character**
 - The semantics associated with a given codepoint
- **Glyph**
 - A particular representation of a character in a given font
- **Character Repertoire**
 - A specific collection of abstract characters; may be open-ended, or closed; may be a subset of all the possible characters in a given encoding.
- **Coded Character Set**
 - A character repertoire connected to an encoding

I Ching: the oldest encoding?

- ~850 BCE? ~2850 BCE?
- 6-bit binary encoding
 - Broken line: Yin
 - Solid line: Yang
- 3-bit chunk defines a “trigram” (eight possible values)
- 64 Hexagrams
 - Each composed of two trigrams
 - Each corresponding to a unique concept

The eight trigrams of the I Ching

	Trigram Figure	Binary Value	Name	Nature	Direction
1	☰	111	Force (乾 <i>qián</i>)	heaven (天)	northwest
2	☱	110	Open (兌 <i>duì</i>)	swamp (澤)	west
3	☲	101	Radiance (離 <i>lí</i>)	fire (火)	south
4	☳	100	Shake (震 <i>zhèn</i>)	thunder (雷)	east
5	☴	011	Ground (巽 <i>xùn</i>)	wind (風)	southeast
6	☵	010	Gorge (坎 <i>kǎn</i>)	water (水)	north
7	☶	001	Bound (艮 <i>gèn</i>)	mountain (山)	northeast
8	☷	000	Field (坤 <i>kūn</i>)	earth (地)	southwest

Cyphers as encodings?

- Codes that use numbers are encodings
 - A=1, B=2, C=3... is an encoding
- Reverse encoding: represent numbers with letters?

Morse Code & the telegraph

- Little used today
- Developed 1830s, first real message sent in 1844
- Morse code defines an encoding and a character set
- Two versions! Fundamental differences...

American/railroad Morse

A	V
B	W
C	X
D	Y
E	Z
F	1
G	2
H	3
I	4
J	5
K	6
L	7
M	8
N	9
O	0
P	,
Q	.
R	!
S	?
T	!
U	!

INTERNATIONAL MORSE CODE

1. A dash is equal to three dots.
2. The space between parts of the same letter is equal to one dot.
3. The space between two letters is equal to three dots.
4. The space between two words is equal to five dots.

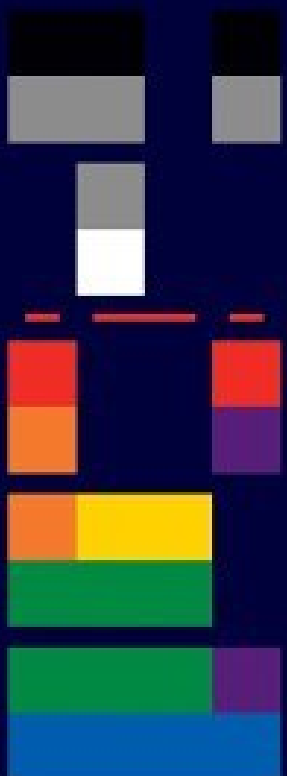
A	• —	U	• • —
B	— • • •	V	• • • —
C	— • — •	W	• — —
D	— • •	X	— • • —
E	•	Y	— • — —
F	• • — •	Z	— — • •
G	— — •		
H	• • • •		
I	• •		
J	• — — —		
K	— • —	1	• — — — —
L	• — • •	2	• • — — —
M	— —	3	• • • — —
N	— •	4	• • • • —
O	— — —	5	• • • • •
P	• — — •	6	— • • • •
Q	— — • •	7	— — • • •
R	• — • •	8	— — — • •
S	• • •	9	— — — — •
T	—	0	— — — — —

Baudot, Murray, and the International Telegraph Alphabet (1874, 1901)

- ITA1, a five-bit encoding and character set, invented by Émile Baudot ca. 1874
 - Five bits = 32 characters, but they also have a shift mechanism for extra characters
- Donald Murray revised ca. 1901
- Further changes by Western Union (and others) created ITA2

The ITA2 Encoding & Character Set

00	01	02	03	04	05	06	07
NUL	E 3	LF	A -	SP	S '	I 8	U 7
08	09	0A	0B	0C	0D	0E	0F
CR	D ENQ	R 4	J BEL	N ,	F !	C :	K <
10	11	12	13	14	15	16	17
T 5	Z +	L >	W 2	H £	Y 6	P 0	Q 1
18	19	1A	1B	1C	1D	1E	1F
O 9	B ?	G &	FIGS	M .	X /	V ;	LTRS
Letters			Figures			Control Chars.	



COLDPLAY X&Y

Baudot, Murray, and the International Telegraph Alphabet (1874, 1901)

- ITA1, a five-bit encoding and character set, invented by Émile Baudot ca. 1874
 - Five bits = 32 characters, but they also have a shift mechanism for extra characters
- Donald Murray revised ca. 1901
- Further changes by Western Union and others created ITA2

The ITA2 encoding and character set

00	01	02	03	04	05	06	07
NUL	E 3	LF	A -	SP	S ' I 8	U 7	
08	09	0A	0B	0C	0D	0E	0F
CR	D ENQ	R 4	J BEL	N , F !	C :	K <	
10	11	12	13	14	15	16	17
T 5	Z +	L >	W 2	H £	Y 6	P 0	Q 1
18	19	1A	1B	1C	1D	1E	1F
O 9	B ?	G &	FIGS	M .	X /	U ;	LTRS
Letters			Figures			Control Chars.	

ASCII (1963)

- American Standard Code for Information Interchange
- ASCII is a 7-bit character set (128 characters, zero to 127)
 - No such thing as “upper ASCII” or “8-bit ASCII”
- 34 control characters plus 94 marking (printable) characters

The printable characters of ASCII

!"#\$%&'()*+,-./
0123456789
:;<=>? @
ABCDEFGHIJKLM
NOPQRSTUVWXYZ
[\]^_`
abcdefghijklmnopqrstuvwxyz
{|}~

EBCDIC (1963-64)

- **Extended Binary Coded Decimal Interchange Code ???**
- 8-bit (single byte) encoding, supports up to 256 characters
- Invented by IBM for use on its mainframes and minicomputers
 - Released with System/360
 - Also used by several other vendors (Fujitsu-Siemens, HP, Unisys)
- Different versions of EBCDIC for different countries
- Double-byte extensions were developed for east Asian countries (CJK)

CCSID 500, an EBCDIC variant

	-0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-A	-B	-C	-D	-E	-F
4-			â	ä	à	á	ã	å	ç	ñ	[.	<	(+	!
5-	&	é	ê	ë	è	í	î	ï	ì	ß]	\$	*)	;	^
6-	-	/	Â	Ä	À	Á	Ã	Å	Ç	Ñ		,	%	_	>	?
7-	ø	É	Ê	Ë	È	Í	Î	Ï	Ì	`	:	#	@	'	=	"
8-	∅	a	b	c	d	e	f	g	h	i	«	»	ð	ý	þ	±
9-	°	j	k	l	m	n	o	p	q	r	ª	º	æ	,	Æ	α
A-	μ	~	s	t	u	v	w	x	y	z	ı	ı	Đ	Ý	Þ	®
B-	¢	£	¥	·	©	§	¶	¼	½	¾	¬		—	ˆ	˜	˘
C-	{	A	B	C	D	E	F	G	H	I		ô	ö	ò	ó	õ
D-	}	J	K	L	M	N	O	P	Q	R	¹	û	ü	ù	ú	ÿ
E-	\	÷	S	T	U	V	W	X	Y	Z	²	Ô	Ö	Ò	Ó	Õ
F-	0	1	2	3	4	5	6	7	8	9	³	Û	Ü	Ù	Ú	

Other single-byte character sets

- IBM/DOS code pages
- Windows code pages (1986)
 - Win-ANSI (codepage 1252) and friends
- MacRoman and other Mac codepages (1985)
- ISO 8859 character sets (1985)
- Many others

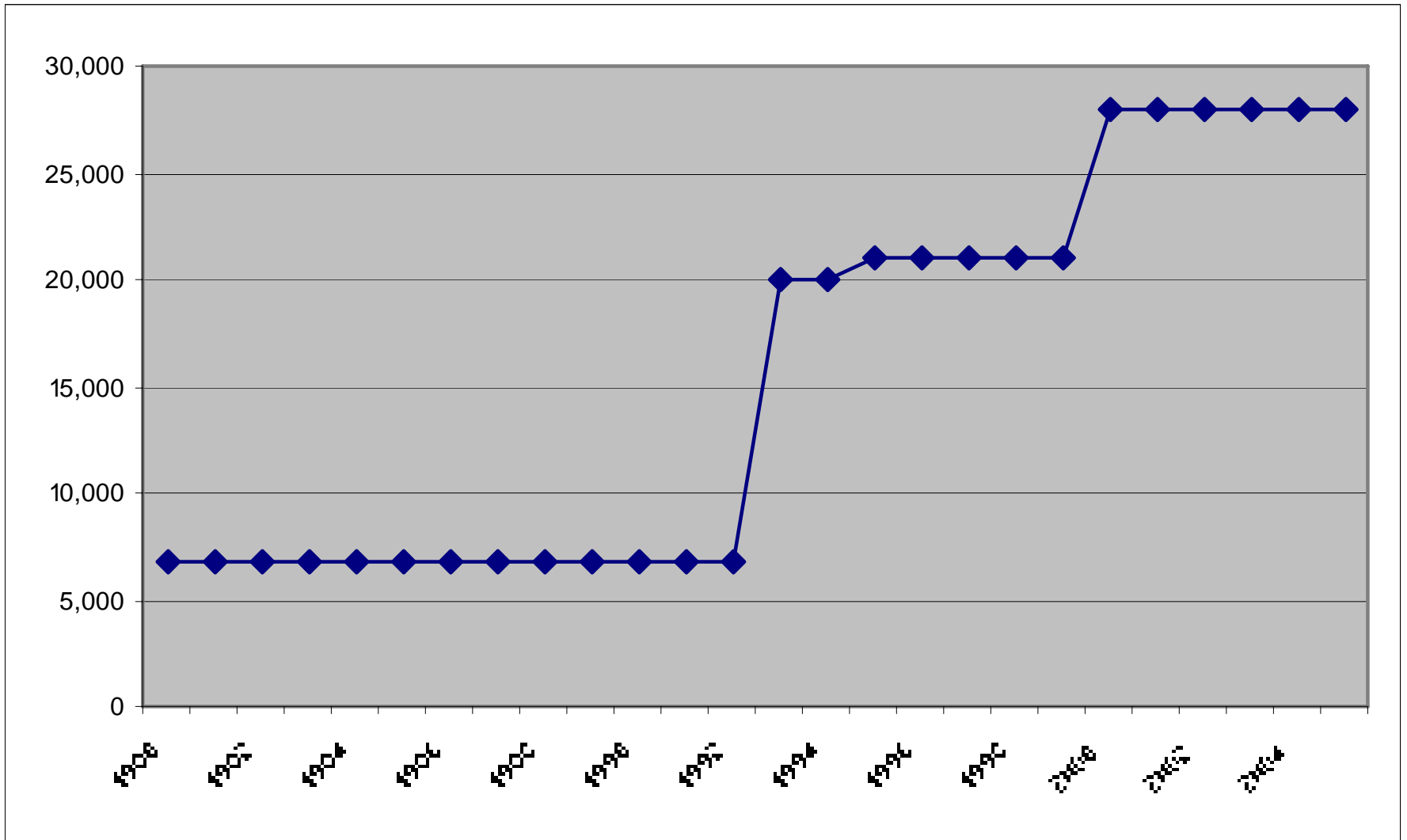
Multi-byte encodings and character sets: Japan

- Japanese Industrial Standards: JIS
 - JIS X 0201 (1997): single-byte, ASCII plus 64 katakana
 - JIS X 0208 (1978, 1997): 6,879 characters
 - **Shift-JIS** (compatible with 0201 and 0208 above)
 - Windows-3.1J / Code page 932
 - Extends Shift-JIS with additional special characters from NEC and IBM
 - JIS X 0213: (2000, 2004) 11,233 characters, including 303 outside the BMP
 - Superset of JIS X 0208

Other multi-byte encodings and character sets: China

- People's Republic of China: *Guójiā Biāozhǔn* (国家标准, GB)
- GB2312-80 (1980)
- GB13000.1-93 (Unicode 1.1, 1993)
- GBK (1993) / Code page 936 (Windows 95)
 - Superset of GB2312, uses same encoding scheme so is backwards compatible
 - Adds characters from GB13000; similar to GB13000 in character set, but different encoding
- GB18030-2000
 - Covers all of Unicode, with defined mapping system.
 - Successor to GBK, backwards compatible
 - GB18030 includes not just Chinese, but “regional languages” such as Korean, Tibetan, Mongolian, Tai-Le, Uyghur and Yi
 - Support for GB18030 is mandatory to sell software in China today!
 - “Support” is very liberally defined

PRC Chinese Character Sets



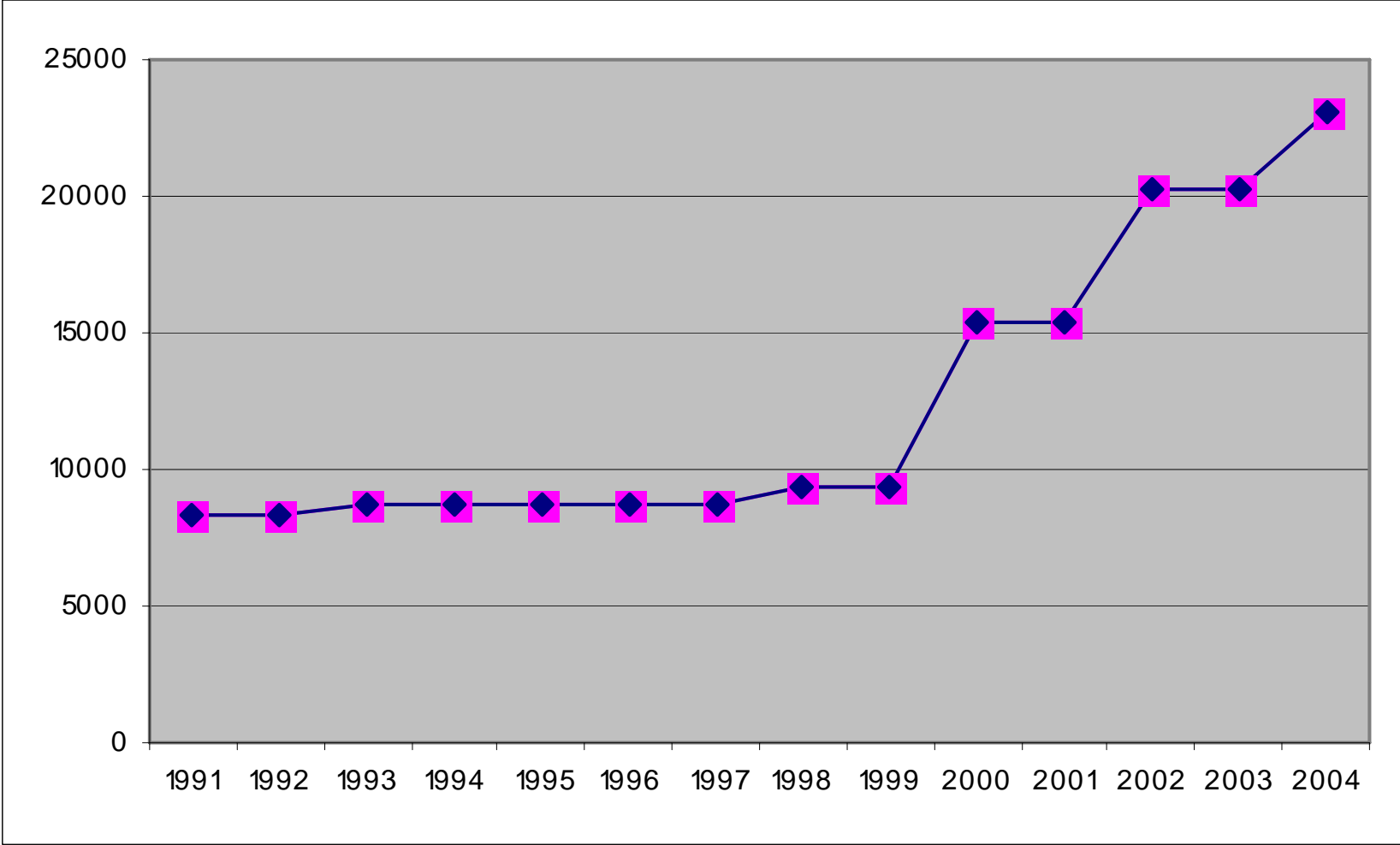
CID “Character Collections”

- Regular Type 1 fonts were “name-keyed”
 - Glyph name is key for encoding
- CID = “Character Identifier” unique numbers used to index and access glyphs
- A “character collection ” is a static glyph set in which each member is identified by a CID
- CID-keyed fonts are supported in both Type 1 and OpenType CFF
 - Apple’s Hiragino Mincho system fonts
- Each character collection is uniquely named:
 - Example: Adobe-Japan1-6
 - /Registry—Its developer, such as Adobe
 - /Ordering—Its purpose, such as Japan1
 - /Supplement—Its incremental additions, such as 6

Evolution of CID Character Collections for Japanese

- Adobe-Japan1-0: 8,284 glyphs, 1991-92
 - Directly compatible with previous OCF format glyph set
- Adobe-Japan1-1: 8,359 glyphs, 1992-93
 - Jis90 (two kanji) and Kanjitalc7 character set
- Adobe-Japan1-2: 8,720 glyphs, 1992-93
 - IBM Selected Kanji character set
- Adobe-Japan1-3: 9,354 glyphs, 1998
 - Pre-rotated glyphs for vertical writing, specifically for OpenType
- Adobe-Japan1-4: 15,444 glyphs, 2000
 - Developed with major Japanese type foundries for commercial/professional publishing
- Adobe-Japan1-5: 20,317 glyphs, 2002
 - Developed in cooperation with Apple; full JIS X 0213 and NLC support
- Adobe-Japan1-6: 23,058 glyphs, 2004
 - Completed JIS X 0212-1990 and U-PRESS support; more JIS78 and JIS X 0213:2004 forms
 - First character set to be compatible with both Win (0212) and Mac OS X (0213) Japanese character sets

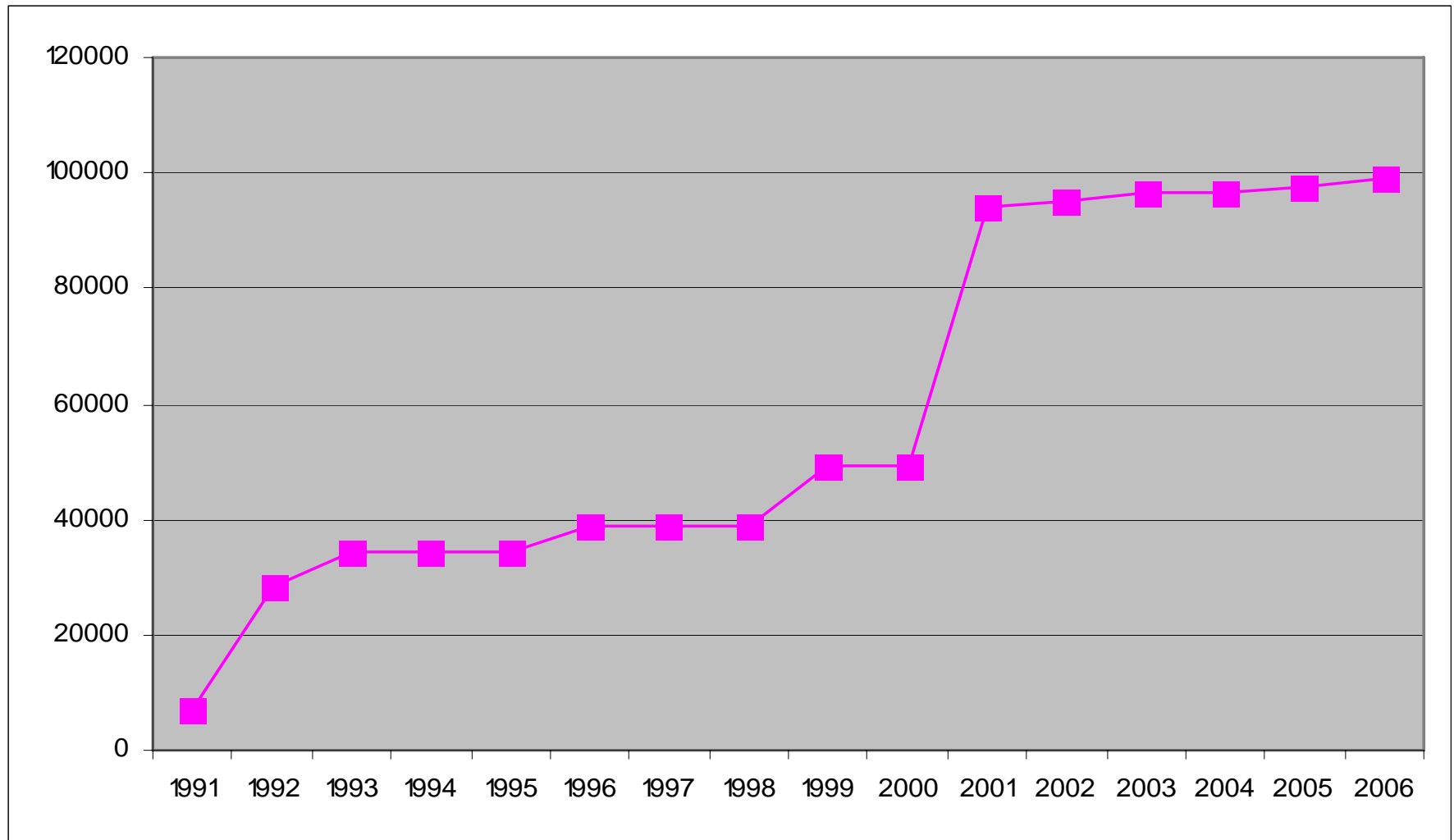
Adobe Japanese CID character set growth



Unicode

- Unique, Universal, Uniform: Unicode
- Variation selectors
- Still relies on predefined character sets and encoding

Encoded characters in Unicode



OpenType

- Access typographic variants and ligatures
 - Beyond the encoded character set
 - All in one font
- Open-ended glyph complement
 - But must have definable relationship to encoded characters

What does it all mean?

- Increased freedom for type designers
 - Design for whatever **languages** you want
 - Design whatever **typographic goodies** you want
 - Not limited to specific defined character sets any more
- Increased end user expectations
 - Users want everything
 - Users all want different things
- Easy for type designers to go over the edge
 - Typical new Adobe western font today has >2000 glyphs

How do we escape endless growth of character/glyph sets?

- There are always more characters/glyphs somewhere that people need/want
- What do we call these unavailable characters or variant glyphs?
- How can we flexibly access them?

Stuff outside the current character set: “gaiji”

- Characters and/or glyphs that are legal in the language, but not in the desired font
- Ideographic (CJK) gaiji
 - Historical variants, personal names, new characters
 - Ideographic writing systems are fundamentally *open-ended*

Other languages have gaiji, too

- Unavailable desired characters arise in all writing systems
- New or local currency symbols
- Add company logo to corporate ID fonts
- Other symbols or new characters

Gaiji examples

辺邊邊邊邊邊邊邊邊邊邊邊邊邊
邊邊邊邊邊邊邊邊邊邊



1913
Personal name "oka"



2004



1997
*euro
currency*



2005
*hryvnia
currency*

SING gaiji solution (2004)

- “Smart Independent Glyphlets”: Single-glyph mini-fonts
- Based on OpenType format
- Can be intended to supplement a particular font, or generic
- Glyph properties can include OpenType layout features
 - Substitution and positioning

How SING works

- *Augmentation*: Glyphlet adds to existing font in memory, without modifying original font on disk
- Glyphlets always *embedded* in document when used—can't lose them

SING Infrastructure Today

- InDesign CS2 supports SING glyphlets for CID-keyed OpenType CFF
 - Chinese, Japanese, Korean OpenType fonts with PostScript style outlines
- Glyphlet creation
 - Illustrator plug-in
 - Coming tools from FontLab

SING Future beyond Creative Suite 3

- *Specifications* already cover TrueType and all OpenType
- Need to extend *implementation* to match
- Boost performance to handle workflows with 10K-100K glyphlets
- Add support to more applications

Conclusion

- Character sets have grown to the breaking point
- New technology offers alternatives to continued growth
- But oversized fonts are here to stay

Better by Adobe.™