Введение в информатику

Е. А. Яревский

физический факультет
СПбГУ
2022

ЛЕКЦИЯ 4

Представление информации в компьютере

Представление символьной информации

60е годы – появление 7 битовой таблицы ASCII
(**A**merican **S**tandard **C**ode for **I**nformation **I**nterchange)
Всего 128 символов:

- 33 кода управляющих символов
- 95 кодов знаков (английский алфавит, цифры, и т.д.)

Только английский!

Русский аналог: КОИ-7 (**К**од для **О**бмена **И**нформацией **7**-битный)
Только русский, естественная транслитерация.

Наследник 7-битовых таблиц – 8-битовые таблицы кодировки.
Фактическая замена: ISO/IEC 8859-1.
Годится для почти всех западноевропейских языков (кроме французского и финского).
Их включение, а также символ евро, привело к появлению в 1999 году ISO/IEC 8859-15.

ASCII Code Chart

	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		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
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

Всего в ISO/IEC 8859 – четырнадцать таблиц.
ISO/IEC 8859-7 – греческий,
ISO/IEC 8859-6 – арабский,
ISO/IEC 8859-5 – русский.

Распространенные/признанные кодировки русских букв

1) ISO/IEC 8859-5 – используется редко(!). Без псевдографики.
2) CP866 (code page) – так называемая альтернативная кодировка, с псевдографикой, MS-DOS. Сейчас используется мало.
3) CP1251 – предложена Microsoft для Windows. Без псевдографики, другие спецсимволы. Очень распространена.
4) koi8-r (CP878) – много используется в Linux, естественная связь с КОИ-7. На базе ГОСТ КОИ-8.
5) CP10007 – предложена Apple, очень похожа на CP1251.

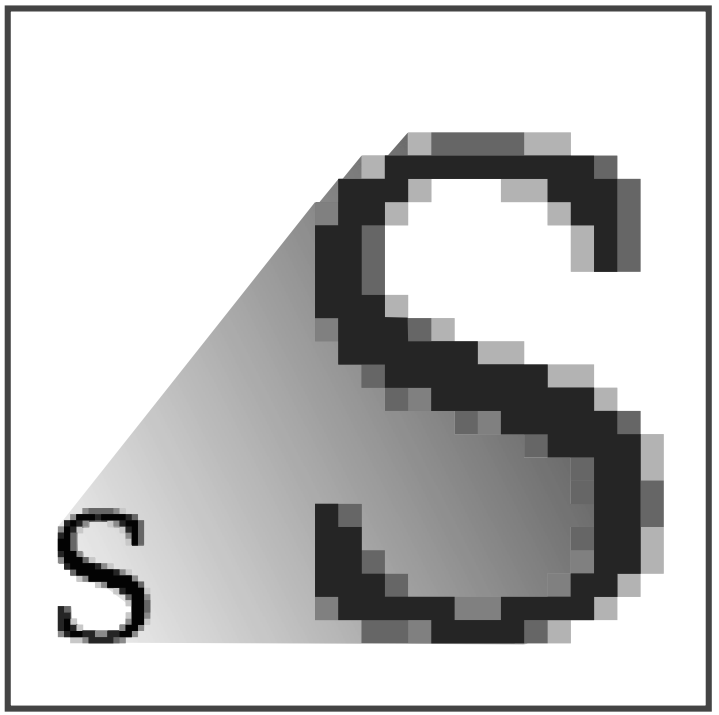
CP1251 (Windows)

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
8	Ђ	Ѓ	‚	ѓ	„	…	†	‡	€	‰	Љ	‹	Њ	Ќ	Ћ	Џ
9	ђ	‘	’	“	”	•	–	—		™	љ	›	њ	ќ	ћ	џ
A		Ў	ў	Ј	¤	Ґ	¦	§	Ё	©	Є	«	¬		®	Ї
B	°	±	І	і	ґ	µ	¶	·	ё	№	є	»	ј	Ѕ	ѕ	ї
C	А	Б	В	Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П
D	Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я
E	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п
F	р	с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я

КОИ 8

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
8	─	│	┌	┐	└	┘	├	┤	┬	┴	┼	▀	▄	▌	▐	▐
9	▒	▒	▓	⌠	■	·	√	≈	≤	≥		⌡	°	²	·	÷
A	═	‖	┌	ё	╓	╔	╕	╖	╗	╘	╙	╚	╛	╜	╝	╞
B	╟	╠	╡	Ё	╢	╣	╤	╥	╦	╧	╨	╩	╪	╫	╬	©
C	ю	а	б	ц	д	е	ф	г	х	и	й	к	л	м	н	о
D	п	я	р	с	т	у	ж	в	ь	ы	з	ш	э	щ	ч	ъ
E	Ю	А	Б	Ц	Д	Е	Ф	Г	Х	И	Й	К	Л	М	Н	О
F	П	Я	Р	С	Т	У	Ж	В	Ь	Ы	З	Ш	Э	Щ	Ч	Ъ

Растровые и векторные шрифты



РАСТР

.jpeg .gif .png

ВЕКТОР

.svg

UNICODE

Проблемы традиционных 8-битовых кодировок
1) Проблема отображения документов в неправильной кодировке.
2) Проблема ограниченности набора символов
3) Проблема преобразования одной кодировки в другую
4) Проблема дублирования шрифтов

Unicode (Юникод) — стандарт кодирования символов, позволяющий представить знаки практически всех письменных языков.

1.1 - 1991-1996 годы (ISO/IEC 10646—1:1993)
16-битовые символы, 2^16 = 65536 символов
2.0 - 1996 год, UTF-16, 2^20+2^16 - 2048 (1 112 064) символов
……
6.2 - 2012 год (кол-во письменностей 100, кол-во символов 110 182)
……
15.0 - 2022 год (кол-во письменностей 161, кол-во символов 149 186)

Стандарт состоит из двух основных разделов:
1) универсальный набор символов (UCS, universal character set).
Задаёт однозначное соответствие символов кодам — элементам кодового пространства, представляющим неотрицательные целые числа (UCS-2, UCS-4).
2) семейство кодировок (UTF, Unicode transformation format).
Определяет машинное представление последовательности кодов UCS.

Изначально 16 бит, обозначают 4мя 16-ричными цифрами

коды от U+0000 до U+007F содержит символы набора ASCII
символы кириллицы: коды от U+0400 до U+052F, от U+2DE0 до U+2DFF, от U+A640 до U+A69F

U+0410 А CYRILLIC CAPITAL LETTER A
U+0462 Ѣ CYRILLIC CAPITAL LETTER YAT
U+046D ѭ CYRILLIC SMALL LETTER IOTIFIED YUS

Различные типы представления UTF

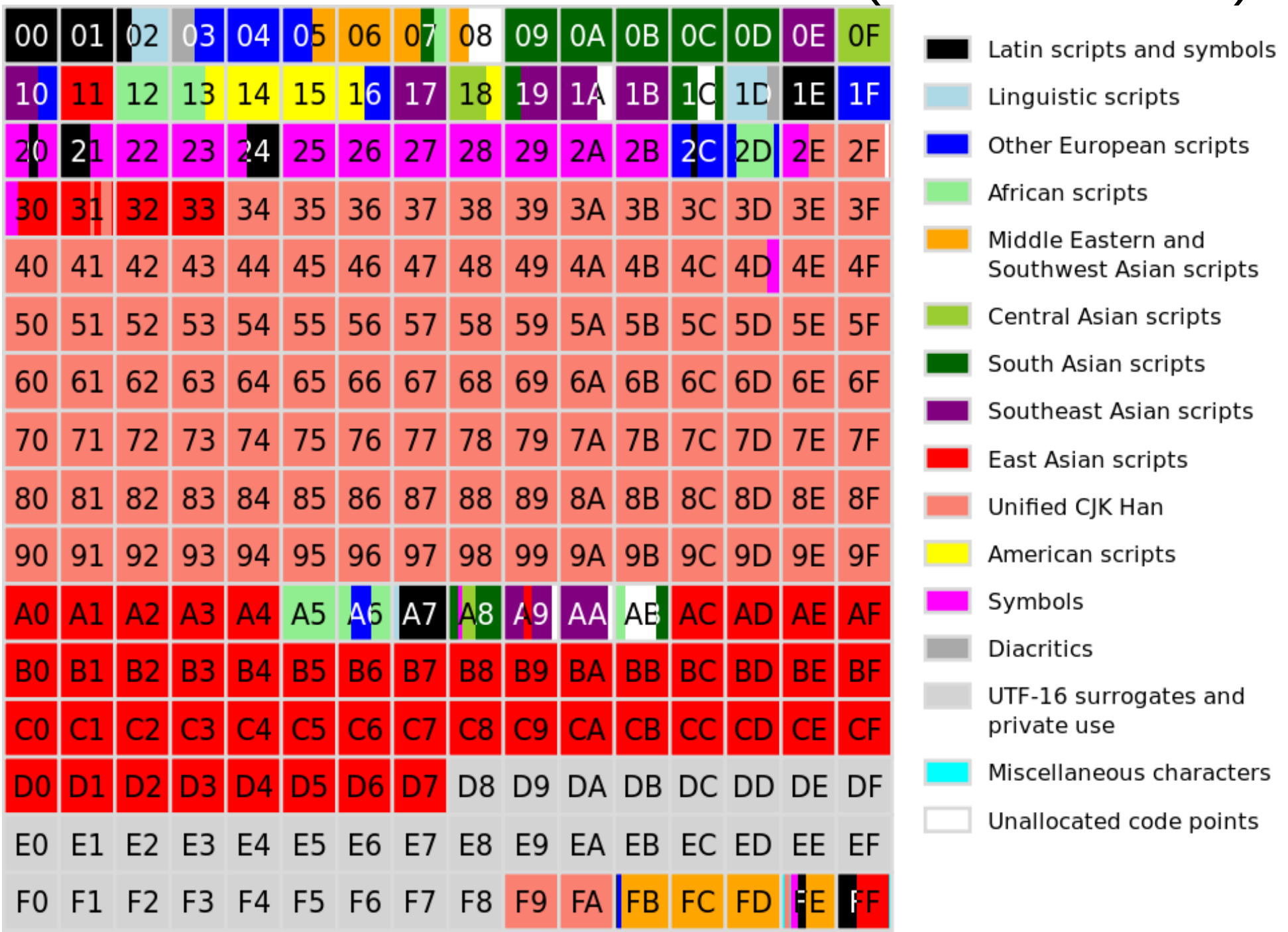
UTF-7 - для передачи по 7-битным каналам, не совместима с ASCII, не включена в стандарт

UTF-8 – обеспечивает наилучшую совместимость с 8-битными символами. 1 байт со значением до 128 – символ ASCII (0ххх хххх). Остальные символы – последовательности от 2 до 6 (фактически, 4) байт. Первый байт имеет вид 11хх хххх, остальные — 10хх хххх.

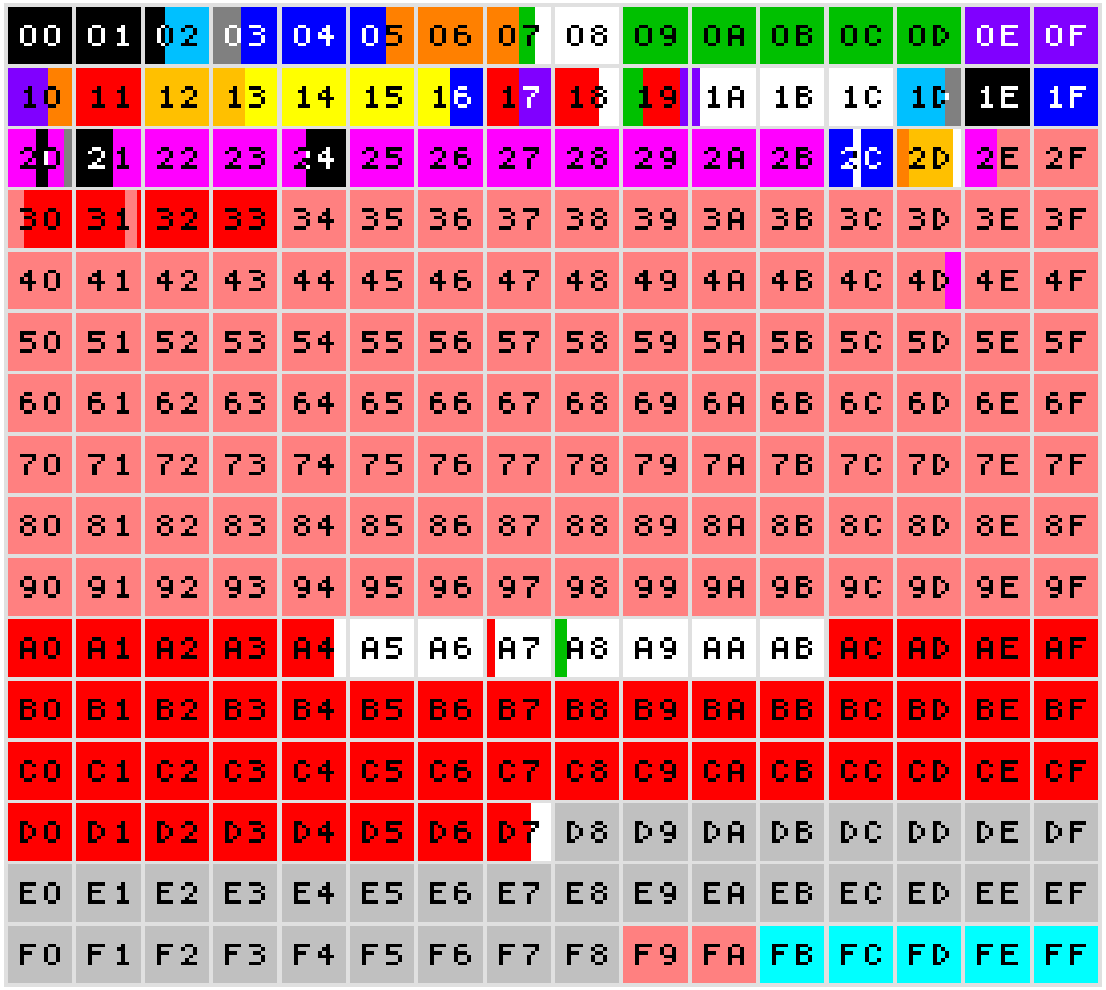
UTF-16 (UTF-16BE, UTF-16LE) – двухбайтовые слова, произвольные значения.
17 плоскостей: от 00 0000 до 10 FFFF (65 536 символов каждая).
U+0000...U+D7FF и U+E000...U+FFFF – обычные символы.
Исключенный диапазон U+D800...U+DFFF – **суррогатные пары**.
Кодируют символы U+010000...U+10FFFF парой 16 битных слов, по 10 бит в слове.
Всего 2^20+2^16 - 2048 = 1 112 064 символов. 2 или 4 байта.

UTF-32 (UTF-32BE, UTF-32LE) – представление 32 битами, 2^32 символов, избыточно.

Базовая плоскость UNICODE (2^{16} символов)



Базовая плоскость UNICODE (2^{16} символов)



Чёрный — расширенный латинский алфав
Голубой — лингвистические символы международного фонетического алфавита IPA;
Синий — другие европейские алфавиты;
Оранжевый — письменности Ближнего Востока;
Светло-оранжевый — письменности Афри
Зелёный — письменности Южной Азии;
Фиолетовый — письменности ЮВА;
Красный — письменности Восточной Азии
Розовый — унифицированные китайско-японско-корейские символы;
Жёлтый — письменности аборигенов Северной Америки;
Пурпурный — символы;
Тёмно-серый — диакритики;
Светло-серый — суррогатные пары UTF-16 и области для частного использования;
Циан — другие знаки;
Белый — не используется.

ЛИТЕРАТУРА

1) Haralambous Y. *Fonts and encodings: from Unicode to advanced typography* (OReilly, 2007), глава 1, 2.
2) Д.М. Хэррис и С.Л. Хэррис, *Цифровая схемотехника и архитектура компьютера* (Morgan Kaufman, 2-е издание, 2013).