ГАПОУ МО «Мурманский колледж экономики и информационных технологий»

Развитие технологии одирования символьной информации в современных условиях

Девиз: «Таблиц кодировок много не бывает»

Автор: Багмет Е.Н.

г. Мурманск,
2016

- Цель исследования:

  Указать основные причины, приведшие к расширению используемых повсеместно кодовых таблиц до стандарта Unicode

- Объекты исследования:

  Кодовые страницы ASCII, Unicode

Обмен информацией

В процессе обмена информации по схемам:

- **человек – человек;**
- **человек – компьютер;**
- **компьютер – компьютер**

представление **информации** происходит в различных формах

Суть набор символов

Набор символов рассматривается, как таблица, задающая кодировку конечного множества символов алфавита (обычно элементов текста: букв, цифр, знаков препинания)

Символы в компьютере обычно кодируются одним или нескольким байтами и объединяются в **кодировки**

Код ASCII

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

Latin-1

	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9	.A	.B	.C	.D	.E	.F
8.	PAD 80	HOP 81	BPH 82	NBH 83	IND 84	NEL 85	SSA 86	ESA 87	HTS 88	HTJ 89	VTS 8A	PLD 8B	PLU 8C	RI 8D	SS2 8E	SS3 8F
9.	DCS 90	PU1 91	PU2 92	STS 93	CCH 94	MW 95	SPA 96	EPA 97	SOS 98	SGCI 99	SCI 9A	CSI 9B	ST 9C	OSC 9D	PM 9E	APC 9F
A.	A0	¡ A1	¢ A2	£ A3	¤ A4	¥ A5	¦ A6	§ A7	¨ A8	© A9	ª AA	« AB	¬ AC	- AD	® AE	¯ AF
B.	° B0	± B1	² B2	³ B3	´ B4	µ B5	¶ B6	· B7	¸ B8	¹ B9	º BA	» BB	¼ BC	½ BD	¾ BE	¿ BF
C.	À C0	Á C1	Â C2	Ã C3	Ä C4	Å C5	Æ C6	Ç C7	È C8	É C9	Ê CA	Ë CB	Ì CC	Í CD	Î CE	Ï CF
D.	Ð D0	Ñ D1	Ò D2	Ó D3	Ô D4	Õ D5	Ö D6	× D7	Ø D8	Ù D9	Ú DA	Û DB	Ü DC	Ý DD	Þ DE	ß DF
E.	à E0	á E1	â E2	ã E3	ä E4	å E5	æ E6	ç E7	è E8	é E9	ê EA	ë EB	ì EC	í ED	î EE	ï EF
F.	ð F0	ñ F1	ò F2	ó F3	ô F4	õ F5	ö F6	÷ F7	ø F8	ù F9	ú FA	û FB	ü FC	ý FD	þ FE	ÿ FF

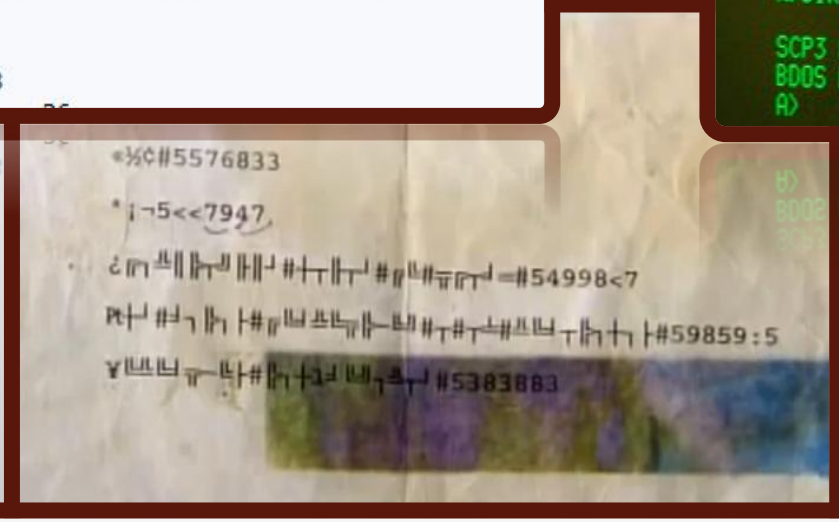
КОИ-8 – русская кодовая таблица

80	81	82	83	84	85	86	87	88	89	8A	8B	8C	8D	8E	8F
90	91	92	93	94	95	96	97	98	99	9A	9B	9C	9D	9E	9F
A0	A1	A2	ё A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD	AE	AF
№ B0	B1	B2	Ё B3	B4	B5	B6	B7	B8	B9	BA	BB	BC	BD	BE	BF
ю C0	а C1	б C2	ц C3	д C4	е C5	ф C6	г C7	х C8	и C9	й CA	к CB	л CC	м CD	н CE	о CF
п D0	я D1	р D2	с D3	т D4	у D5	ж D6	в D7	ь D8	ы D9	з DA	ш DB	э DC	щ DD	ч DE	ъ DF
Ю E0	А E1	Б E2	Ц E3	Д E4	Е E5	Ф E6	Г E7	Х E8	И E9	Й EA	К EB	Л EC	М ED	Н EE	О EF
П F0	Я F1	Р F2	С F3	Т F4	У F5	Ж F6	В F7	Ь F8	Ы F9	З FA	Ш FB	Э FC	Щ FD	Ч FE	Ъ FF

Разнообразие кодировок вызвало невозможность удобного обмена информацией

Заголовок страницы (title)...
Исходный заголовок:
p p p p p p p p p c c p p рёр рё p p p p p p p c c p p p p p p p p p p рёс p p p p p c p c рёр p c c p
Обработанный заголовок (без стоп-слов):
p p p p p p p p p p p рёр рё p p p p p p p p p p p p p p p p p рёс p p p p p p рёр p p

Стоп-слов в заголовке: 8



Р В§Р ВµР В"Р С•Р Р†Р ВµР С" РЎРґР ВµР в"-РЎвЂЎР В°РЎРґ РЎС"Р Р†Р С'Р Т'Р С'РЎвЂљ Р В"Р С'РЎв,¬РЎРЉ РЎвЂљР С•, РЎвЂЎРЎвЂљР С• Р С•Р В¶Р С'Р Т'Р В°Р ВµРЎвЂљ РЎС"Р Р†Р С'Р Т'Р ВµРЎвЂљРЎРЉ.

Небольшой размер ASCII (256 символов), а также аналогичных таблиц не позволял закодировать большое количество символов национальных алфавитов

В настоящее время распространение получил новый стандарт Unicode, который отводит на каждый символ не 1 байт, а 2 байта

Юникод, как система для линейного представления текста

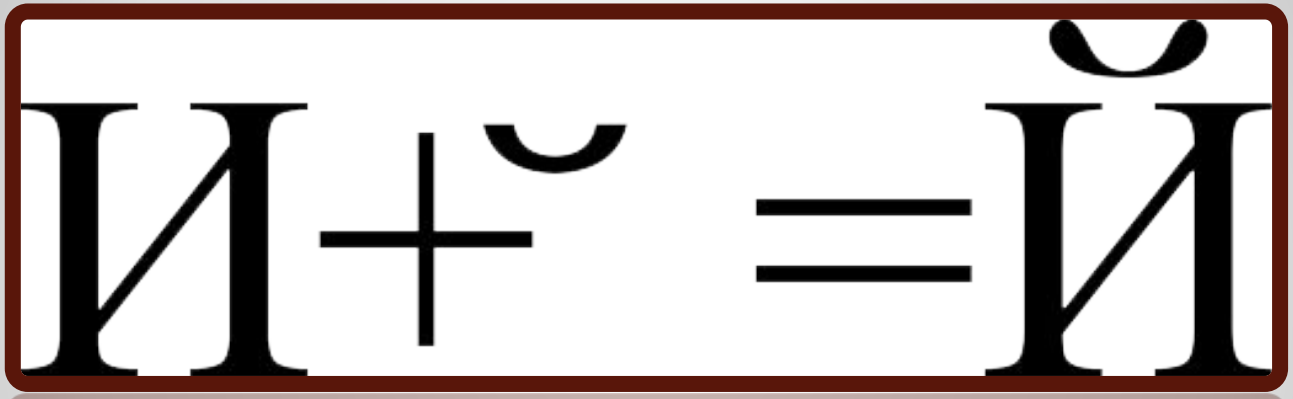
Символы, имеющие дополнительные над- или подстрочные элементы, могут быть представлены в виде построенной по определённым правилам последовательности кодов (составной вариант, composite character) или в виде единого символа (монолитный вариант, precomposed character).

На 2014г. считается, что все буквы крупных письменностей в Юникод внесены, и если символ доступен в составном варианте, дублировать его в монолитном виде не нужно.

Общее количество символов в Unicode равно 65536 (2^16=1024*64). Из них отдано:

- 336 – на латынь;
- 256 – на русский язык;
- по 128 – на некоторые редкие национальные алфавиты

Отдельное кодирование диакритических символов для экономии места



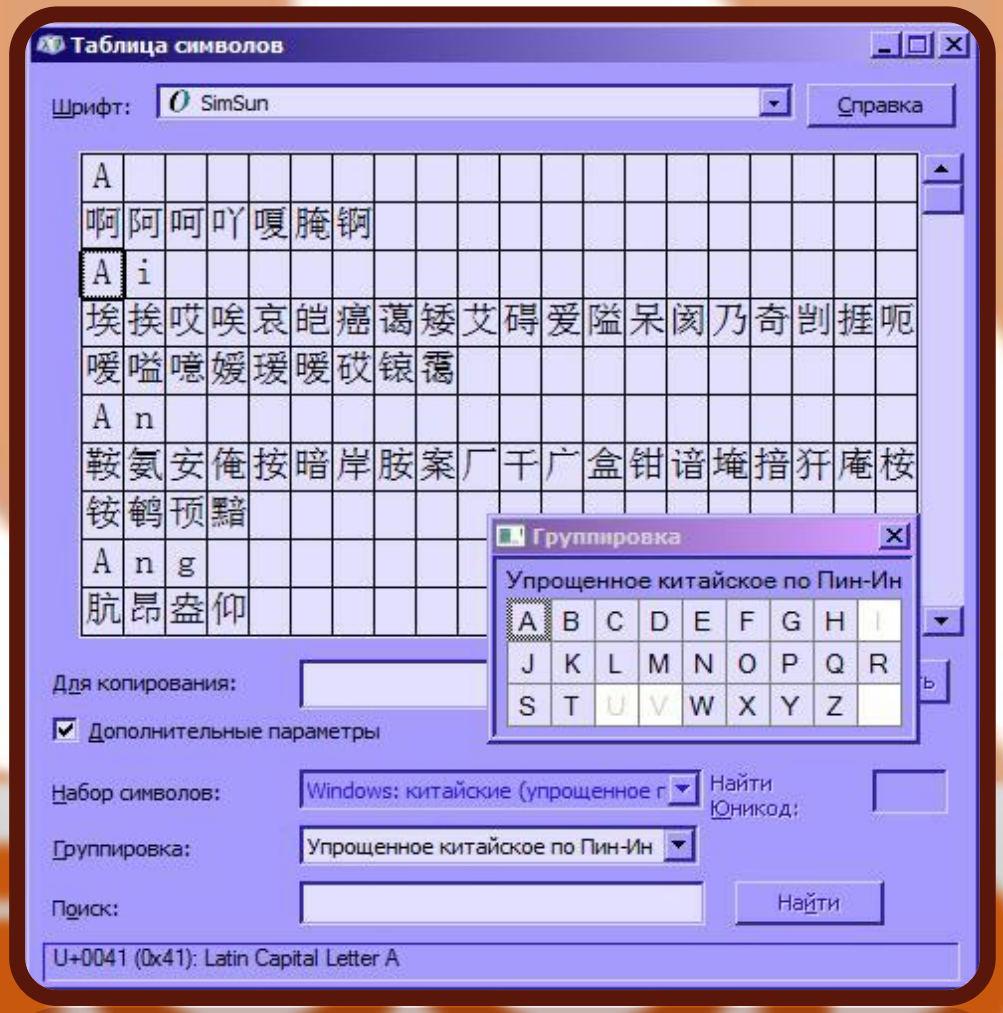
Представление символа «Й» в виде базового символа «И» и модифицирующего символа «˘»

Кодировки русского алфавита и Unicode (ISO 10646)

0x0	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
400		Ё	Ђ	Ѓ	Є	S	І	Ї	Ј	Љ	Њ	Ћ	Ќ		Ў	Џ
410	А	Б	В	Г	Д	Е	Ж	З	И	Й	К	Л	М	Н	О	П
420	Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я
430	а	б	в	г	д	е	ж	з	и	й	к	л	м	н	о	п
440	р	с	т	у	ф	х	ц	ч	ш	щ	ъ	ы	ь	э	ю	я
450		ё	ђ	ѓ	є	s	і	ї	ј	љ	њ	ћ	ќ		ў	џ
490	Ґ	ґ	Ғ	ғ			Җ	җ			Қ	қ				
4A0			Ң	ң											Ү	ү
4B0	Ұ	ұ	Ҳ	ҳ							Һ	һ				
4D0									Ә	ә						
4E0									Ө	ө						

Вспомогательная таблица для ручного преобразования кодировок символов

Китайские иероглифы – 20992 позиций



Универсальная система кодирования (**Юникод**) представляет собой набор графических символов и способ их кодирования для компьютерной обработки текстовых данных

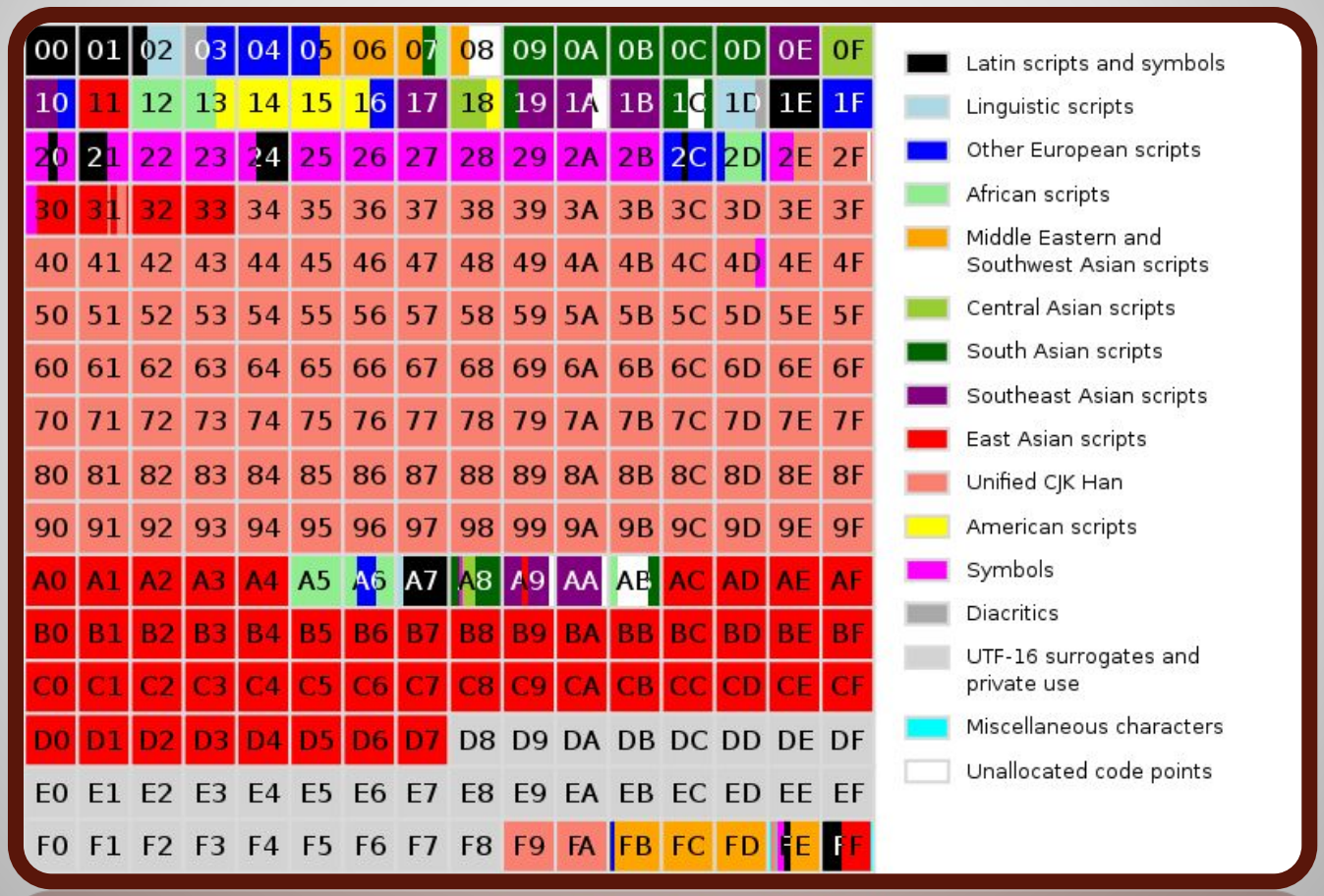
Графические символы - это символы, имеющие видимое изображение. Графическим символам противопоставляются управляющие символы и символы форматирования.

Графические символы

Графические символы включают в себя следующие группы:

- буквы, содержащиеся хотя бы в одном из обслуживаемых алфавитов;
- цифры;
- знаки пунктуации;
- специальные знаки (математические, технические, идеограммы и пр.);
- Разделители.

Распределение представленных в Unicode символов



Проблемы Unicode

Не решенными остаются такие вопросы:

- Не предусмотрено переключение горизонтального и вертикального написания для японского и китайского языков;
- Перевод из строчных букв в заглавные зависит от языка;
- До сих пор не всё прикладное программное обеспечение поддерживает корректную работу с ним.

Вывод

Таким образом:

- Unicode стал настоящим спасением, заменив собой все разнообразие локальных кодировок;
- Несмотря на это, Unicode все же не смог разрешить все проблемы, связанные с интернационализацией кодирования.

Список литературы

1. American National Standards Institute http://www.ansi.org
2. UTF-8, a transformation format of ISO 10646 http://tools.ietf.org/html/rfc2279
3. Internet Assigned Numbers Authority – организация, отвечающая за административное управление в Интернете – http://www.iana.org


Спасибо за внимание!