

CAP 2. DADOS MULTIMÍDIA

AULA 7: REPRESENTAÇÃO DE CARACTERES

Cap. 2 Dados Multimídia

Conteúdo

- Processo de captura de áudios, imagens e vídeos
- Representação digital de áudios, imagens e vídeos
- Representação de caracteres/textos
- Principais características e requisitos das informações multimídia

Nesta aula veremos...

- Representação de caracteres/textos







Representação de Caracteres

Caracteres

- **Multimídia:** Mídia são as portadoras de informação (formas de transmitir alguma informação)
- Palavras e símbolos, falados ou escritos, são a forma mais comum de comunicação
 - Meio adequado para transmitir informações essenciais de modo preciso
 - Forma principal de comunicação assíncrona (defasado no tempo), e quase tempo-real (mensagens instantâneas) entre pessoas

Representação de Caracteres

- **Natureza dupla dos textos**

- **Conteúdo léxico**, é a parte do texto que transmite o seu significado (sua semântica)
 - Caracteres abstratos: não importa a aparência dos caracteres para o entendimento da semântica
- **Aparência**, atributos visuais dos caracteres (fonte, tamanho, disposição na tela, etc.)
 - A representação visual de um caractere denomina-se Glifo.
 - Caractere abstrato “A” pode ter uma infinidade de representações gráficas, incluindo “”, “”, “”, “”, “”, “”.

Representação de Caracteres

Formas possíveis do texto

- Texto não formatado (plain text)
 - número de caracteres disponíveis é limitado
 - representação simples (dimensão dos caracteres é fixa e não permite diferentes fontes ou estilos)
- Texto formatado (rich text)
 - aparência mais rica, várias fontes, cores, estilos e dimensões
 - produzidos por processadores de texto
- Hipertexto
 - texto ao qual se adicionam hiperligações originando texto não linear
 - permite navegação entre documentos de texto.

Representação de Caracteres

Caracteres abstratos

- São os caracteres representados apenas quanto a sua natureza léxica:
 - São agrupados em alfabetos;
 - Cada idioma ou grupo de idiomas usa um alfabeto.



Conjuntos de caracteres

- São tabelas mantidas pelo sistema operacional que consistem em uma correspondência entre os códigos e os caracteres
- Contém representações de grafemas (unidades fundamentais de um sistema de escrita) ou unidades similares a grafemas
 - Incluem maiúsculas, minúsculas, sinais de pontuação, números e símbolos matemáticos.

Representação de Caracteres

Vantagens da utilização de conjuntos de caracteres:

- É vital guardar os caracteres na forma de códigos:
 - Para tornar o texto revisável (não imagem) e permitir a busca;
 - Para facilitar a comparação de caracteres (basta comparar códigos)
- Permitem associar os caracteres dos teclados a representação desses caracteres:
 - Por exemplo, quando se pressiona um A no teclado, esse caractere é procurado na tabela de caracteres para depois ser apresentado no monitor.

Normalização é o mais importante

- Pois os códigos universais podem facilmente ser trocados entre máquinas diferentes e que usam sistemas operacionais diferentes.

Representação de Caracteres

ASCII - American Standard Code for Information Interchange

- Primeiro conjunto de caracteres normalizado (1968)
- Adequado à língua inglesa
 - Usa 7 bits para representar cada código: 128 (2^7) caracteres no total
- Insuficiente para muitas línguas (128 caracteres é limitado)

<i>Bits</i> 3210	654 <i>000</i>	<i>001</i>	<i>010</i>	<i>011</i>	<i>100</i>	<i>101</i>	<i>110</i>	<i>111</i>
0000	NUL	DLE	SP	0	@	P	\	p
0001	SOH	DC1	!	1	A	Q	a	q
0010	STX	DC2	"	2	B	R	b	r
0011	ETX	DC3	#	3	C	S	c	s
0100	EOT	DC4	\$	4	D	T	d	t
0101	ENQ	NAK	%	5	E	U	e	u
0110	ACK	SYN	&	6	F	V	f	v
0111	BEL	ETB	'	7	G	W	g	w
1000	BS	CAN	(8	H	X	h	x
1001	HT	EM)	9	I	Y	I	y
1010	LF	SUB	*	:	J	Z	j	z
1011	VT	ESC	+	;	K	[k	{
1100	FF	FS	'	<	L	\	l	
1101	CR	GS	-	=	M]	m	}
1110	SO	RS	.	>	N	^	n	~
1111	SI	US	/	?	O	_	o	DEL

Exemplo:

A = `_100 0001` (41h – 65d)

V = `_101 0110` (56h – 96d)

Representação de Caracteres

ISO 8859

- Normaliza os conjuntos de caracteres de 8 bits (10 partes):
 - ISO 8859-1: ISO Latin1, caracteres utilizados na maioria dos países da Europa Ocidental, primeiros 128 caracteres são os mesmos do ASCII de 7 bits, os restantes 128 são códigos para os idiomas europeus
 - ISO 8859-2: ISO Latin2, para outros idiomas da Europa Oriental (Checo, Eslovaco, Croata)
 - ISO 8859-5: Cirílico
 - ISO 8859-7: Grego moderno
 - ISO 8859-8: Hebreu

Representação de Caracteres

ISO 8859-1

128	Ç	144	É	160	á	176	☒	193	⊥	209	〒	225	ß	241	±
129	ü	145	æ	161	í	177	☒	194	⌞	210	⌞	226	Γ	242	≥
130	é	146	Æ	162	ó	178	☒	195	⌞	211	ℒ	227	π	243	≤
131	â	147	ô	163	ú	179		196	—	212	ℓ	228	Σ	244	∫
132	ä	148	ö	164	ñ	180	⌞	197	⊕	213	ℓ	229	σ	245	∫
133	à	149	ò	165	Ñ	181	⌞	198	⌞	214	ℓ	230	μ	246	÷
134	â	150	û	166	ª	182	⌞	199	⌞	215	⌞	231	τ	247	≈
135	ç	151	ù	167	º	183	⌞	200	ℒ	216	⌞	232	Φ	248	°
136	ê	152	—	168	¿	184	⌞	201	ℓ	217	⌞	233	Θ	249	.
137	ë	153	Ö	169	—	185	⌞	202	ℒ	218	⌞	234	Ω	250	.
138	è	154	Ü	170	¬	186	⌞	203	〒	219	■	235	δ	251	√
139	ï	156	£	171	½	187	⌞	204	⌞	220	■	236	∞	252	—
140	î	157	¥	172	¾	188	⌞	205	=	221	■	237	φ	253	²
141	ì	158	—	173	¡	189	⌞	206	⌞	222	■	238	ε	254	■
142	Ä	159	ƒ	174	«	190	⌞	207	⊥	223	■	239	∩	255	
143	Å	192	Ł	175	»	191	⌞	208	ℒ	224	α	240	≡		

Representação de Caracteres

A opção pelas variantes ISO 8859 acaba por não conseguir resolver bem o problema:

- 7+1 bits são claramente insuficientes para representar todas as línguas (Chinês, japonês etc.)
- E os textos multilíngue? Como se trabalha com várias línguas simultaneamente?

Representação de Caracteres

Unicode



- Consórcio de empresas (Adobe, Apple, Microsoft, ...) definiram Unicode
 - As linguagens HTML, XML e Java usam o Unicode.
- Padrão que permite aos computadores representar e manipular, de forma consistente, texto de qualquer sistema de escrita existente
 - Desenvolvido em conjunto com um Conjunto Universal de Caracteres (UCS – Universal Character Set – ISO/IEC 10646), que contém mais de 128000 caracteres abstratos, cada um identificado por um nome não ambíguo a um número inteiro (code point)

Representação de Caracteres

Unicode

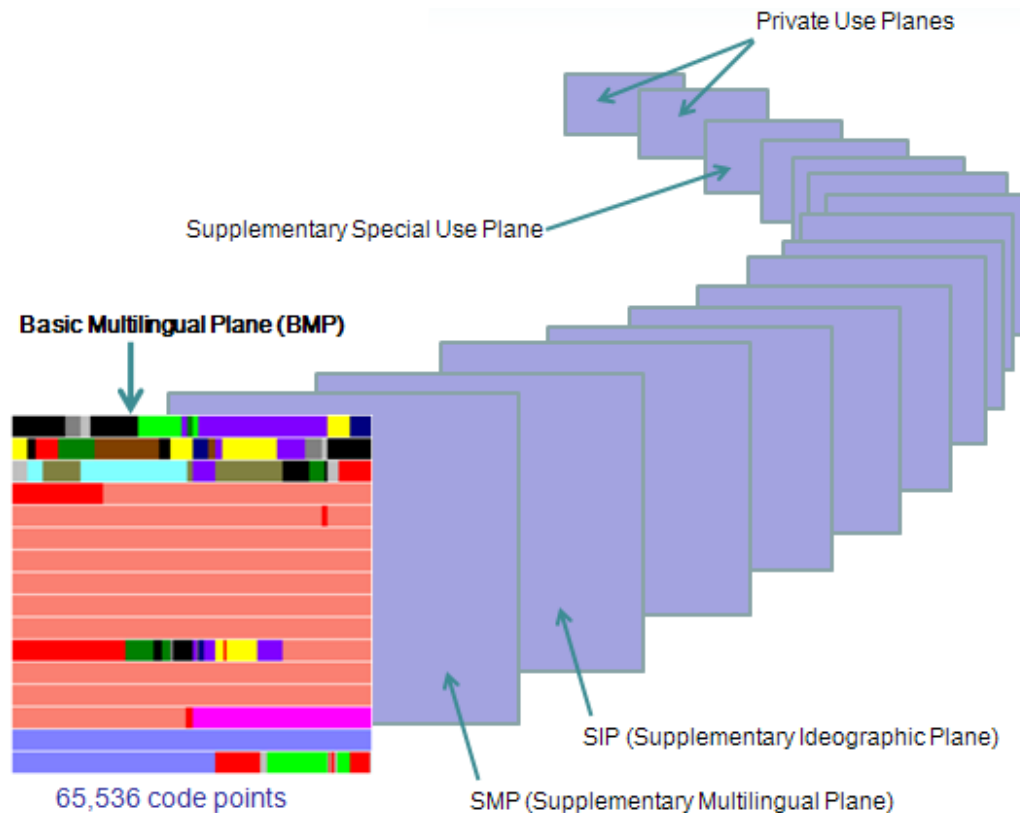


- Consiste de
 - um repertório de mais que 128000 caracteres cobrindo 100 scripts (coleção de letras e outros signos escritos usado para representar uma informação textual em um ou mais sistemas de escritas),
 - uma metodologia para codificação
 - uma enumeração de propriedades de caracteres (como caixa alta e caixa baixa)
 - um conjunto de arquivos de computador com dados de referência
 - Regras para normalização, decomposição, ordenação alfabética e renderização.

Representação de Caracteres

Unicode

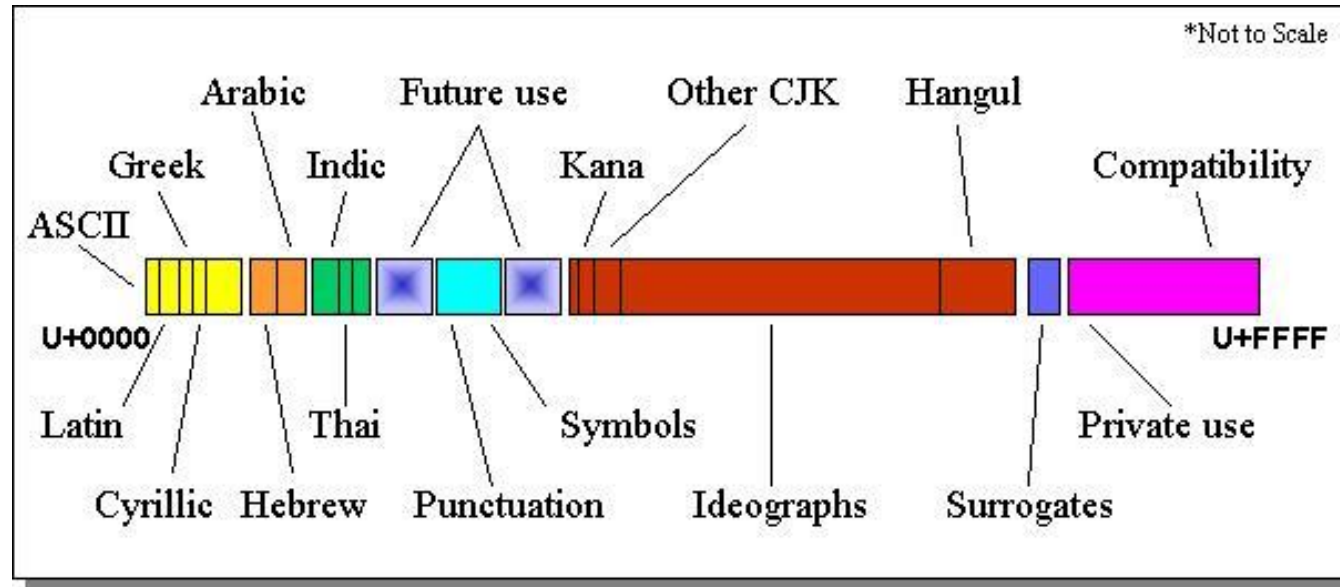
- Padrão Unicode codifica caracteres em um espaço numérico entre 0 a 10FFFF
- Espaço de codificação é dividida em 17 planos (numerados de 0 a 16)



Representação de Caracteres

Unicode

- Layout de codificação Unicode do BMP (Plano 0)



- Exemplo codificação de caractere:
 - LETRA MAIÚSCULA LATINA A, U+0041.
 - U+aaaa é um valor de código: U+ se refere a valores de código Unicode, e aaaa representa um número de quatro dígitos hexadecimais de um caractere codificado.

Representação de Caracteres

Unicode

- Padrão Unicode codifica caracteres em um espaço numérico entre 0 a 10FFFF
- Existem alguns formatos de codificação destes valores
 - UTF-8, UTF-16 e UTF-32.
- UTF-8
 - uma forma de codificação de tamanho variável, requer de um a quatro bytes para expressar cada caractere Unicode
 - "A" é 41 (mesmo que no ASCII!)
 - α é CE 91
 - Katakana "A" é E3 82 A2 ア
 - Gothic Ahsa é F0 90 8C B0 Ა

Representação de Caracteres

- **Fontes e faces**

- **Face** é uma família de caracteres que normalmente inclui muitos tamanhos e estilos de tipos
 - Arial, Times New Roman e Courier New são exemplos de faces
- **Fonte** é um conjunto de caracteres de um único tamanho e estilo pertencente a uma família de face particular.
 - *Times 15 pontos itálico* é uma fonte
 - As fontes digitais são versões das fontes tradicionais (algumas do século XV)
 - As fontes podem ser vistas como tabelas de correspondência entre os caracteres abstratos e a sua representação gráfica (grifo)

Representação de Caracteres

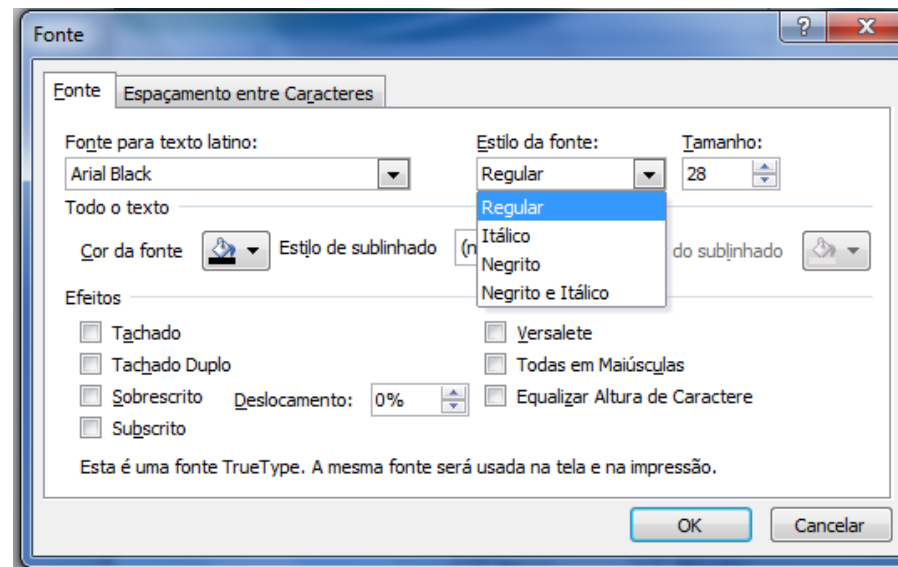
Fontes

- Duas possibilidades de armazenamento
 - Armazenados em arquivos e instalados no sistema operacional:
 - Compartilhados por todos os arquivos e todas as aplicações
 - Quanto são requeridas e não existem tem de ser trocadas por fontes alternativas
 - São embutidas nos próprios arquivos:
 - Vantagem importante para o designer de uma aplicação multimídia pois é livre de usar qualquer fonte no seu trabalho.
 - Não se compartilham as fontes entre documentos que usam as mesmas fontes.

Representação de Caracteres

Tamanhos e estilos

- Tamanhos geralmente são expressos em pontos;
 - um ponto corresponde a 0,0138 polegadas ou aproximadamente 1/72 de uma polegada.
- Os estilos normais das fontes são negrito, itálico (oblíquo) e sublinhado
 - outros atributos como contorno de caracteres podem ser adicionados pelo programa.



Pontos Importantes

Mapas de caracteres

- Mantém a relação de códigos representando caracteres

Padrões de codificação

- ASCII 7 bits: inicial para o inglês
- ISO8859 (8 bits):
 - várias partes para conjuntos de idiomas diferentes,
 - não suficiente para representar vários idiomas
 - não permite multilíngue
- Unicode solução mais adotada hoje