Technology Leadership Council
Brazil

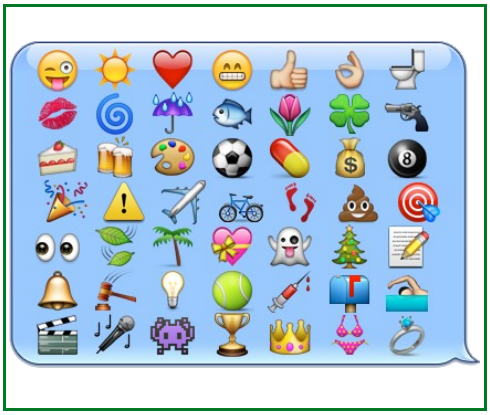
Mini Paper Series Ano 7
Agosto, 2012 - Nº 164

Unicode ♥ וניקוד ☻ Уникод ♫ يونيكود

Avi Alkalay

Você sabia que há pouco tempo era impossível misturar diversas línguas numa mesma frase de texto sem a ajuda de um editor multilíngue especial? Mais ainda, que havia idiomas cujas letras sequer tinham uma representação digital sendo impossível usá-las em computadores? Tudo isso virou passado com o advento do Unicode e para entendê-lo vamos relembrar alguns conceitos:

• **caractere** – É a representação digital do que chamamos de letra ou grafema ou ideograma. Alguns exemplos de caracteres: J (jota maiúsculo), ç (cê-cedilha minúsculo), Φ ζ λ Ψ Ω π (caracteres gregos), וניקוד (a palavra "Unicode" em hebraico), símbolos matemáticos como × ÷ ∞ ∂ ∑ ∫, ou financeiros como $ ¢ £ ¥ ₪ €, hieróglifos egípcios e muitos outros que mostraremos neste texto;



• **glifo** – Uma representação gráfica para um certo caractere. As fontes Times New Roman e Arial usam glifos diferentes para representar o mesmo caractere "g";

• ***encoding*** – É uma dica que damos ao computador para que ele saiba qual caractere ou letra humana ele deve usar para mostrar certo código binário. Por exemplo, o código 224 no encoding ISO-8859-1 é o caractere "à", porém no ISO-8859-8 é a letra "א". Repare que no universo desses antigos *encodings* as letras "à" e "א" não podem coexistir porque usam o mesmo código binário. E é justamente esse o problema que foi levantado no começo do texto.

Antes do Unicode era necessário somente 1 *byte* de computador para armazenar a informação de 1 caractere. Os *encodings* são necessários mas, como vimos, havia limitações indesejáveis. Como o Unicode propõe uma gama muito maior de códigos binários, único e imutável por ideograma, caracteres de idiomas diferentes podem agora coexistir no mesmo texto. Neste exemplo "à" e "א" tem códigos Unicode que não conflitam entre si: 0x00ED e 0x05D0.

A história do Unicode começa em 1987 na Xerox e Apple, e tenta incorporar todos os ideogramas e letras do mundo, um conjunto bem maior do que 255 caracteres (que é o que cabe em 1 *byte*). Um caractere Unicode pode ter de 1 a 4 *bytes*.

Evoluir para múltiplos *bytes* por caractere tem certas implicações pois os softwares não estavam preparados para isso. Contar caracteres numa frase é agora diferente de contar o número de *bytes* ocupados por essa frase. Mostrar ou imprimir tal frase é também agora uma tarefa de outra ordem: há línguas onde se escreve da direita para esquerda, como árabe ou hebraico, versus as da esquerda para direita, baseadas no sistema latino. No título do artigo há a palavra "Unicode" em ambos os sentidos na mesma frase, na escrita latina (→), hebraica (←), russa (→) e árabe (←) respectivamente e isso serve de exemplo para mostrar que a questão de múltiplos sentidos de escrita na mesma frase é contemplada e resolvida pelo Unicode.

O Unicode introduziu também desafios de desempenho pois há muito mais caracteres maiúsculos e minúsculos para comparar e mais *bytes* para armazenar e processar. Mas tudo isso é marginal com a evolução do poder computacional, universalidade e eternidade da informação que o Unicode oferece.

Ainda no título, outra coisa que chama a atenção são símbolos como ♪♠☼☺, ideogramas que fazem parte de uma faixa de caracteres do Unicode chamada Emoji, incorporado ao padrão em 2010. Mas por enquanto, podemos utilizar somente alguns Emojis em forma de texto porque estão em fase de implementação nos sistemas operacionais. Por outro lado, eles já são bastante populares nos sistemas iOS (iPhone, iPad), Mac OS X Lion e Linux. Somente a versão 8 do Windows terá suporte completo a Emoji.

Emoji é também um marco de evolução da linguagem escrita. Em tempos de uso intenso de redes sociais e SMSs, é muito mais divertido e expressivo escrever "eu ♥ você", "estou com fome², vamos❣", "★adorei ☺☺", "hoje estou zen ◒" etc. E que tal mais esses para seu próximo tweet?: ↗ ☠ ☢ ☭ ☣ ✡ ✝ ➡ ☮ ☎ ♔ ♛ ✿. Todos são caracteres tão comuns quanto "ú" ou "H" e graças ao Unicode, não é necessário nenhum recurso de processador de texto para usá-los.

O Unicode já está em pleno uso na Internet. É comum encontrarmos páginas que misturam línguas ou usam caracteres avançados. Um relatório periódico do Google mostra que entre 2008 e 2012 o uso de Unicode em sites subiu de 35% para mais de 60%. Não poderia ser menos, pois Unicode é uma tecnologia absolutamente essencial para um mundo globalizado e multiucultural.

Ao longo deste texto mostrei alguns caracteres, letras e ideogramas curiosos. Para fechar, deixo-os com uma última ideia: um pouco de unicode não faz mal a ninguém.

Para saber mais

http://www.coisinha.com.br/viravira/

http://www.DecodeUnicode.org/

http://en.wikipedia.org/wiki/Emoji

http://googleblog.blogspot.com.br/2012/02/unicode-over-60-percent-of-web.html

Avi Alkalay é Arquiteto de Informação, com 20 anos de experiência em tecnologia de informação, formado em 1995 pela UNESP e membro do TLC-BR desde 2006. O **Mini Paper Series** é uma publicação quinzenal do **TLC-BR** e para assinar e receber eletronicamente as futuras edições, envie um e-mail para **tlcbr@br.ibm.com**. O **TLC-BR** é a representação da IBM Academy of Technology no Brasil.