

B / Unicode, UTF

Introduction à Unicode

Beyrouth, le 24 mai 2005

Patrick Andries
patrick@hapax.qc.ca

ASCII

	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	CAL	EM	SUB	ESC	FS	GS	RS	US
2	sp	!	"	#	\$	%	&	'	(	)	*	-	,	-	.	/
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
8																
9																
A																
B																
C																
D																
E																
F																

ISO 8859-1/Latin-1

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0																
1																
2	sp	!	"	#	\$	%	&	'	(	)	*	+	,	-	.	/
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
8																
9																
A	VRSP	;	€	£	¤	¥	¦	§	¨	©	ª	«	¬	®	¯	
B	°	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿
C	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	
D	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	
E	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	
F	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	

24 mai 2005

Tutoriel Unicode, Beyrouth

7

JIS X 0201

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOI	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAL	EM	SUB	ESC	FS	GS	RS	US
2	sp	!	"	#	\$	%	&	'	(	)	*	-	,	-	.	/
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
8																
9																
A	sp	。	「	」	、	・	ヲ	ア	イ	ウ	エ	オ	ヤ	ユ	ヨ	ツ
B	ー	ア	イ	ウ	エ	オ	カ	キ	ク	ケ	コ	サ	シ	ス	セ	ソ
C	タ	チ	ツ	テ	ト	ナ	ニ	ヌ	ネ	ノ	ハ	ヒ	フ	ヘ	ホ	マ
D	ミ	ム	メ	モ	ヤ	ユ	ヨ	ラ	リ	ル	レ	ロ	ワ	ン	・	・
E																
F																

24 mai 2005

Tutoriel Unicode, Beyrouth

8

EBCDIC

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NIL				FF	HT	LC	DEL								
1				TM	RES	NL	BS	IL								
2	DS	SOS	FS		BP	LF	EOB	PRE								
3					PN	RS	OC	EOT								
4	sp										¢	.	<	(	+	
5	&										!	\$	*	)	;	~
6	-	/										,	%	_	>	?
7											:	#	@	'	=	"
8		a	b	c	d	e	f	g	h	i						
9		j	k	l	m	n	o	p	q	r						
A			s	t	u	v	w	x	y	z						
B																
C		A	B	C	D	E	F	G	H	I						
D		J	K	L	M	N	O	P	Q	R						
E			S	T	U	V	W	X	Y	Z						
F	0	1	2	3	4	5	6	7	8	9						

Mojibaké

- Que se passe-t-il quand on envoie une suite de caractères dans un code et on le lit avec un autre code ?

- Cafouillis

Les Japonais ont un mot pour ce concept

- 文字化け, *moji* caractère + *baké* changé
- « 文字化け » pourra ainsi être affiché
« •¶Žš%o», - »

Pourquoi Unicode ?

- Résoudre le problème du mojibaké
 - Affectation d'un seul numéro à chaque caractère
 - Codage simple (éviter les problèmes de copier-coller avec des codes multi-octets à états)
- Permettre à un même logiciel de prendre en charge un max. d'écritures (moins de versions)

Vous avez-dit Unicode?

*Unicode est un standard qui définit un
numéro unique pour chaque caractère,
quelle que soit la plate-forme,
quel que soit le logiciel,
quelle que soit la langue.*

Unicode en unicode

- يونيكود
- 유니코드
- ூନିକୋଡ
- 統一碼
- Unicode
- 統一碼
- ynikod
- Γιοὺνικοντ
- Юникод
- யூனிகோட்
- ூନିକୋড
- யூனிகోড
- ூନିକୋড

24 mai 2005

Tutoriel Unicode, Beyrouth

14

Omniprésent

- Produits qui mettent en œuvre en partie ou complètement Unicode :
 - Windows NT, Windows 2000, Windows XP
 - Java, .NET
 - Mac OS 9, OS X, Linux (en partie)
 - Oracle, Sybase, ...
 - XML, XSL, XHTML, HTML etc.

24 mai 2005

Tutoriel Unicode, Beyrouth

15

ISO/CEI 10646 et Unicode

Même répertoire :

- même caractères
- mêmes noms
- mêmes numéros
- Résultat d'une fusion des deux normes légèrement différentes à l'origine

ISO/CEI 10646 et Unicode

- Unicode publié en anglais uniquement (bien qu'une traduction annotée en français existe sur <<http://hapax.qc.ca>>)
- Unicode définit plus de comportements (formes normalisées, coupures de lignes, algorithme bidi)
- Unicode fournit plus d'information sur les caractères et écritures (catégorie, casse, etc.)

ISO/CEI 10646 et Unicode

Membres :

- SC2
 - Organismes nationaux (membres votants)
 - AFNOR, SNIMA, DIN sont membres du JTC1/SC2/GT2
 - Membres observateurs et de liaison (consortium Unicode)
- Unicode
 - sociétés (Microsoft, Apple, Sun, Sybase, Adobe,...)
 - associations (libraires, sociétés bibliques)
 - particuliers
 - rares organismes nationaux également dans SC2

Dix principes de conception

- 10 principes qui ont guidé élaboration
- Tous les principes ne s'appliquent pas toujours de manière égale
- Adhésion à un principe entraîne parfois violation à d'autres
- Unicode et ISO/CEI 10646 sont le résultat d'un délicat équilibre

1. Universalité

On peut à la limite écrire dans un seul texte

Dvořák 天竺 سلام α ψ ∩ \$ ¥

→ Pas besoin de plusieurs versions de logiciel pour chacune de ces écritures

2. Efficacité

- Tous les caractères sont d'une égale facilité d'accès
- Par convention les caractères sont regroupés en un bloc de caractères consécutifs
 - Compression possible des données appartenant principalement à un bloc
- Partage des caractères de ponctuation

3. Caractères et non glyphes

- b, ب et ب sont tous des U+0062 ب
- Unicode ne s'intéresse qu'au codage (affectation d'un numéro aux caractères)
- Pas au rendu (affichage, impression)
- Pour certaines écritures (celles de l'Inde, l'arabe), glyphes nécessaires nettement plus nombreux que caractères

3. Caractères et non glyphes

ب
لبنان بيروت

3. Caractères et non glyphes

f f i

ffi

ffi

3. Caractères et non glyphes

سلام

கோ

4. Sémantique

- casse du caractère : majuscule Å, minuscule a, casse de titre Dz
- directionnalité : gauche, droite, faible, neutre
- à symétrie miroir : $< \int \rightarrow$ (à ajuster selon directionnalité du texte)
- numéral : iv 7
- combinatoire : ٠ ١ ٢ ٣

5. Texte brut

- Pas de balisage, pas de formatage
- Texte enrichi ou texte de fantaisie
 - Texte brut + balisage/formatage

Brut : un texte brut

Enrichi : `<h1>Un texte <b>enrichi</b></h1>`

5. Texte brut et texte enrichi

Fais ce que dois, advienne que pourra

Fais ce que dois, advienne que pourra

Fais ce que dois, advienne que pourra

FAIS CE QUE DOIS, ADVIENNE QUE POURRA

Fais ce que dois, advienne que pourra

Fa**i**s ce que d**o**is, ad**v**ienne que p**o**urra

6. Ordre logique

En mémoire

L i b a n

U+004D U+0061 U+0072 U+006F U+0063

ﻝ ﻭ ﻥ ﻟ ﻭ ﻝ ﻭ ﻥ ﻟ

U+0721 U+0307 U+0720 U+071F U+0307 U+0710

À l’affichage

Liban

فلافل

- Utilise les propriétés de directionnalité d’Unicode

7. Unification

- *a* de *chat*, *Katze*, *gato* et *kat* codé une seule fois (unification par écriture et non langues)
- Unicode code donc les écritures, pas les sons
- On unifie aussi, bien sûr, les glyphes pour ne coder qu'un caractère

7. Unification

B

- 212C MAJUSCULE B DE RONDE
 - = fonction de Bernoulli

B ≠ <anglaise> B

≈ <police> B

7. Unification han (CJCV)

Pourquoi ?

- Similaire au cas des lettres latines unifiées (il existe des lettres espagnoles ou allemandes)
- CJCV hérité du chinois (comme le latin en Europe)
- Impossible de reproduire fidèlement chacune [37!] des normes de base (énorme ensemble de caractères identiques !)

7. Unification han (CJCV)

Comment ?

- Unifie des jeux de caractères, pas des langues
 - Chinois, japonais, coréen (et vieux vietnamien)
- Séparation des sources
 - Pas d'unification au sein d'un jeu-source
- Beaucoup d'unifications évidentes
 - Long processus par équipes d'experts (GRI) pour les cas les plus délicats

7. Unification – non unifiés

- Aspect similaire, signification différente

土

571F

terre

士

5E8B

lettré

- Pas d'unification au sein d'un jeu-source

區

5340

zone, quartier, secteur

区

533A

8. Composition dynamique

- Composition dynamique
 - $e + \circ \rightarrow \hat{e}$
- Processus ouvert, très productif
- Le $\circ$ peut donc s'appliquer à des voyelles ou des consonnes dans des langues qui utilisent l'écriture latine ou non

8. Composition dynamique

$e + \hat{\circ} + \grave{\circ} = e + \grave{\circ} + \hat{\circ} = \hat{e}$

- Diacritique ou signe combinatoire
- Suit la lettre de base (e)
- Placement centrifuge
- Par convention, nombre infini de diacritiques peuvent s'adjoindre à un caractère de base

8. Caractères de base et diacritiques

Caractères *Glyphe*

e ö ó ù ı ➔ ě

Empilement de diacritiques

प + ि ➔ पि

Signe vocalique indien antéposé (i bref)

8. Composition dynamique (arabe)

مهندس

مهندس

• م + ُ + ه + َ + ن + ُ + د + ِ + س

9. Séquence équivalente

LJ + U

≡

L + J + U

A + ُ + ِ

≡

A + ِ + ُ

A + ُ + َ

≠

A + َ + ُ

E + ِ + ِ

≠

E + ِ + ِ

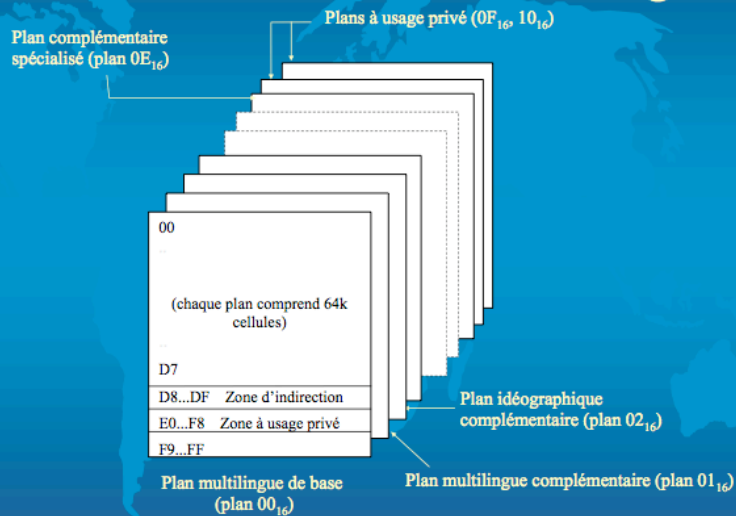
10. Convertibilité

- Il a fallu définir au début des méthodes
 - Importation
 - Exportation
- entre Unicode ISO/CEI 10646 et les jeux de caractères fondamentaux préexistants
- et garantir des conversions aller-retour sans perte
- S'oppose parfois au principe d'unification

Plans de codage

- l'espace de codage est divisé en plans de 64K cellules
- 1 cellule = max 1 caractère

Plans et zones de codage



24 mai 2005

Tutoriel Unicode, Beyrouth

50

PMB

- Latin (étendu, p. ex. API)
- Grec, copte, cyrillique
- Arménien, géorgien
- Arabe, syriaque, hébreu
- Écritures de l'Inde
- Thaï, lao, tibétain, birman
- Chinois, japonais, coréen (CJC)
- Ponctuation, flèches, symboles math.

24 mai 2005

Tutoriel Unicode, Beyrouth

51

PMC

- PMB surtout écritures vivantes
- PMC, écritures mortes et symboles supplémentaires:
 - Linéaire B
 - Italique (ancien italien)
 - Gotique (pas gothique)
 - Ougaritique
 - Symboles musicaux
 - Symboles alphanumériques mathématiques (ω , $\mathfrak{B}$, $\mathbb{B}$, $\mathfrak{M}$)

Lire les tableaux Unicode

Code	Glyphe	Description	
00AE	®	SYMBOLE MARQUE DÉPOSÉE	<i>nom ISO 10646</i>
00AF	—	MACRON	<i>nom ISO 10646</i>
		= barre supérieure APL	<i>nom optionnel</i>
		• ce caractère chasse	<i>renseignement</i>
		→ 02c9 ~ lettre modificative macron	<i>renvois</i>
		→ 0304 ~ diacritique macron	
		→ 0305 ~ diacritique tiret haut	
		≈ 0020 0304 ~	<i>décomposition de compatibilité</i>

Lire les tableaux (suite)

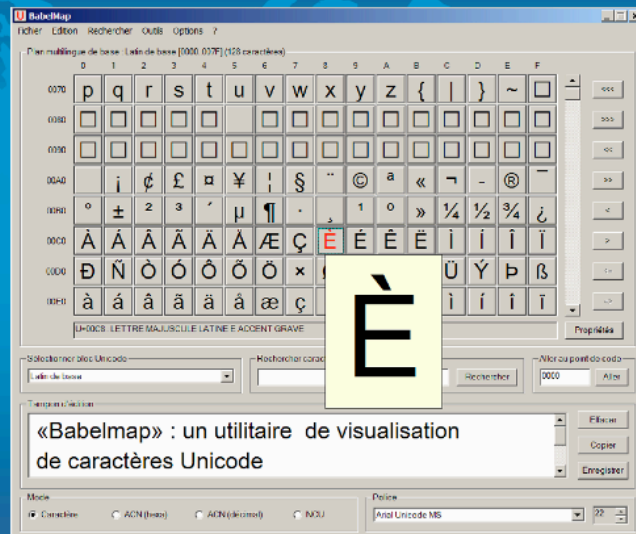
Code	Glyphe	Description
00E5	å	LETTRE MINUSCULE LATINE A ROND EN CHEF
		• danois, norvégien, suédois, wallon
		≡ 0061 a 030A °

décomposition canonique

Propriétés de caractères

- Peuvent être utilisées dans des algorithmes de
 - Rendu
 - Découpe en ligne
 - Tri
 - ...
- Propriétés Unicode
 - Catégorie générale
 - Classe bidi (directionnalité de la lettre)
 - Décomposition
 - Classes combinatoires canoniques...
 - ...

Babelmap

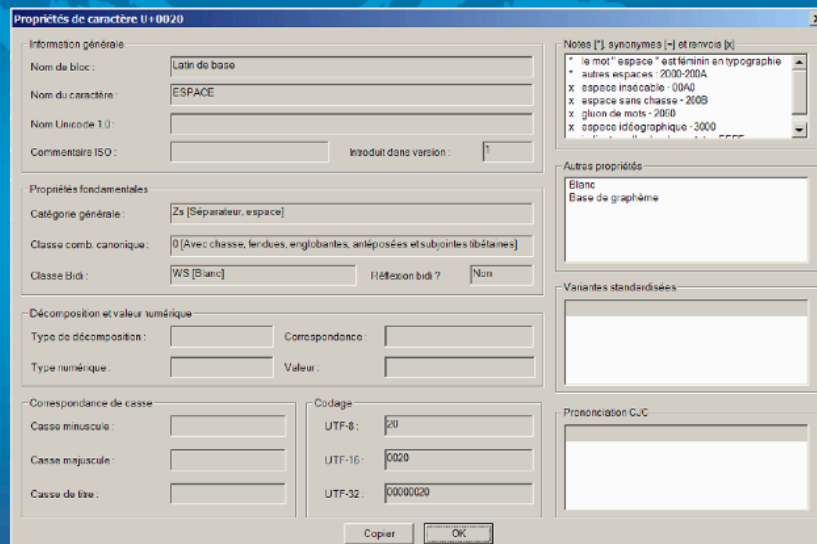


24 mai 2005

Tutoriel Unicode, Beyrouth

58

Propriétés



Catégorie générale

- Lettre, majuscule
- Lettre, minuscule
- Lettre, modificateur
- Marque, à chasse nulle
- Nombre, chiffre décimal
- Ponctuation,
- Symbole, mathématique
- Symbole, devise monétaire
- Etc.

Modèle de codage des caractères

- Forme que prendront les numéros de caractères (codets) lors d'un stockage ou d'un transfert.
- Unicode définit un modèle de codage de caractères à 4 niveaux de représentation des caractères :
 1. Répertoire de caractères abstraits
 2. Jeu de caractères codés
 3. Forme en mémoire des caractères
 4. Mécanisme de sérialisation de caractères

Répertoire

- Ensemble de caractères abstraits, habituellement d'un alphabet connu
- Abstrait car définis par convention (voir les 26 lettres de l'alphabet latin)
- Comprend des caractères, et non des glyphes.
- Ensemble non ordonné.
- Unicode a un répertoire ouvert, contrairement à la plupart des jeux de caractères existants

Jeu de caractères codés

- Correspondance entre un ensemble de caractères abstraits et un ensemble d'entiers non négatifs.
- Ce dernier ensemble peut ne pas être contigu.
- On dit qu'un caractère abstrait est codé dans un jeu de caractères donné si un numéro de caractère existe pour ce caractère.

Forme en mémoire

- Aussi appelée forme « naturelle » des caractères
- Unités de stockage en mémoire
 - Entier d'une certaine largeur (exemples : octet ou seize) qui sert d'unité de base à l'expression des numéros de caractère dans la mémoire d'un ordinateur.
 - Le nombre d'unités de stockage représentant un caractère est variable.
 - Pour jeux de caractères traditionnels, le plus souvent une seule forme des caractères en mémoire (p.ex. ASCII, Latin-1).

UCS-4 et UTF-32

- UCS-4 = UTF-32
- chaque numéro de caractère est représenté par une quantité sur 32 bits.
- Espace de code est arbitrairement limité à 0..10FFFF pour des raisons de compatibilité avec UTF-16 (voir plus loin)

Exemples UCS-4/UTF-32

- LETTRE MINUSCULE GRECQUE DELTA
 - δ
 - N° de caractère : U+03B4
 - Unité de stockage en UCS-4 : 0x000003B4
- LETTRE GOTIQUE D
 - 𝔻
 - N° de caractère : U+10333
 - Unité de stockage en UCS-4 : 00010333

UTF-16

- 16 bits
- PMB sont codés avec un seul seizeset,
- autres plans codés à l'aide de deux seizesets (dits d'indirection) :
 - un seizeset d'indirection supérieur [D800..DBFF]
 - un seizeset d'indirection inférieur [DC00..DFFF]

Exemples UTF-16

- LETTRE MINUSCULE GRECQUE DELTA
 - δ
 - N° de caractère : U+03B4
 - Unité de stockage en UTF-16 : 0x03B4
- LETTRE GOTIQUE D
 - Ɔ
 - N° de caractère : U+10333
 - Unité de stockage en UTF-16 : 0xD800, 0xDF33

UTF-8

- chaque numéro de caractère est représenté par une suite de 1 à 4 octets.
- Espace de code est arbitrairement limité à 0..10FFFF pour des raisons de compatibilité avec UTF-16 (voir ci-dessus)

Exemples UTF-8

- LETTRE MINUSCULE GRECQUE DELTA
 - δ
 - N° de caractère : U+03B4
 - Unité de stockage en UTF-8 : 0xCE, 0xB4
- LETTRE GOTIQUE D
 - Ɔ
 - N° de caractère : U+10333
 - Unité de stockage en UTF-8 : 0xF0, 0x90, 0x8C, 0xB3

Pourquoi tous ces formats ?

- UTF-8
 - très populaire avec HTML et autres protocoles car ASCII préservé
 - Marche avec logiciel avec peu de changements
- UTF-16
 - Compromis entre accès efficace et place réduite
- UTF-32
 - populaire quand place pas un problème

Décomposition

- Unicode code parfois des caractères sous deux formes :
 - forme précomposée (pour des raisons historiques)
 - forme décomposée (caractère de base + diacritique, plus productive)
- Exemple :
 - U+00C5 (Å)
 - U+0041 (A) U+030A (°)

Équivalence canonique

- Caractères considérés comme identiques (et qui ne diffèrent donc même pas au niveau visuel).
 - É (U+00C9) est une variante canonique de E (U+0045) + ◌ (U+0301)
- De manière formelle :
 - On dit que deux suites de caractères sont des équivalents canoniques si leurs décompositions canoniques complètes (récursives) respectives sont identiques

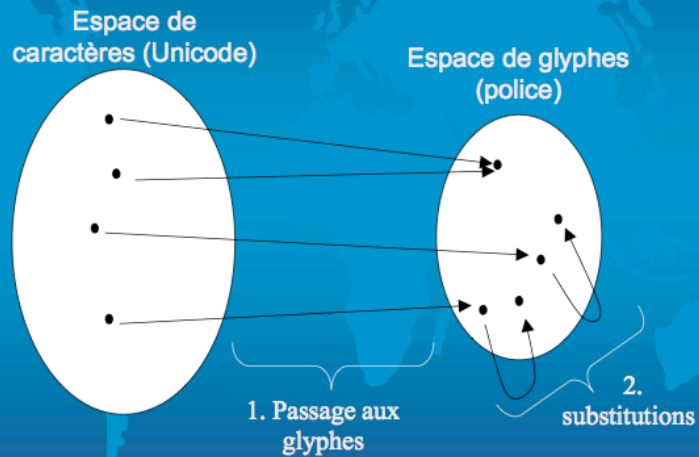
Unicode

- Ne se préoccupe que des caractères.
- Caractères sont stockés en « ordre logique », c'est-à-dire ordre phonétique.
- Unicode suppose la présence d'un **moteur de rendu qui pourra, le cas échéant**, réordonner les caractères et pourra substituer des glyphes et placer de manière précise.

Pourquoi ce modèle ?

- Garantir que les opérations suivantes fonctionneront :
 - Correcteur grammatical, traduction automatique
 - Indexation
 - Recherche et remplacement de texte
 - Copier-coller
 - Exportation de texte
 - Changement de police
 - Changement de style de paragraphe ou de caractères

Sélection des glyphes



24 mai 2005

Tutoriel Unicode, Beyrouth

91

Algorithme bidirectionnel

- En mémoire, ordre logique
- Chaque caractère à une directionnalité implicite
 - Forte : lettres, signes (marques) explicites
 - Faible : chiffres européens et signes numéraux («+», « , » «- », « . »,...)
 - Neutres : autres (blancs, «(», «)», «!», «?», ...)
- Affichage des neutres et faibles dépend du contexte

24 mai 2005

Tutoriel Unicode, Beyrouth

92

Ce qu'Unicode n'est pas



Questions

Si j'ai écrit un document avec Unicode 3.2, est-ce qu'il sera lu exactement de la même façon avec Unicode 4.1 ?

- « Avec » ? Unicode n'est pas un produit...
- Le répertoire est ouvert, mais compatible vers le bas
- On ne supprime jamais un caractère codé, on ne peut qu'en ajouter au répertoire

Questions (suite)

Pour se conformer à Unicode faut-il prendre en charge 95 000 caractères Unicode ?

La conformité peut s'énoncer sommairement de manière informelle de la manière suivante :

- On peut ignorer un caractère qu'on ne comprend pas, mais pas le corrompre
- La mise en œuvre d'un sous-ensemble ne dépend que de vous
- U+0007 : <commande> sonnerie (driing !)

Ressources

- ISO et Unicode en français
 - <http://hapax.qc.ca>
 - Tableaux de caractères
 - Textes explicatifs (plus de 300 pages)
- Livre « Codage et fontes »
de Yannis Haralambous
- Liste de diffusion et forum de discussion
 - (Unicode Afrique)
 - <http://fr.groups.yahoo.com/group/Unicode-Afrique/>
 - (Unicode fr)
 - nntp://fr.comp.normes.unicode