

A world map rendered in a light blue color, centered on the Atlantic Ocean, serving as a background for the text.

# Formes et propriétés Unicode

Patrick Andries

[hapax@iquebec.com](mailto:hapax@iquebec.com)

# Formats et propriétés Unicode

A faint, light blue world map is visible in the background of the slide, centered behind the text.

- Modèle de codage des caractères
  - Formes codées d'Unicode
- Propriétés des caractères
- Formes normalisées

# Modèle de codage des caractères

- Forme que prendront les numéros de caractères lors d'un stockage ou d'un transfert.
- Unicode définit un modèle de codage de caractères à 5 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
  5. Surcodage de transfert

# 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
  - Ꝁ
  - N° de caractère : U+10333
  - Unité de stockage en UCS-4 : 00010333

# UCS-2

- chaque numéro de caractère est représenté par seize bits (un «seizet»)
- $2^{16}$  valeurs d'unités de codage, donc maximum potentiel de  $2^{16}$  numéros de caractère
- cette forme n'existe qu'en ISO/CEI 10646
- elle ne permet que d'adresser les caractères du PMB

# Exemple UCS-2

- LETTRE MINUSCULE GRECQUE DELTA
  - $\delta$
  - N° de caractère : U+03B4
  - Unité de stockage en UCS-2 : 0x03B4
- LETTRE GOTIQUE D
  - $\mathfrak{D}$
  - N° de caractère : U+10333
  - Unité de stockage en UCS-2 : **inaccessible**

# 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]

# Indirection UTF-16

- Caractères complémentaires :
  - [0xD800-0xDBFF] = 0x400 = 1024 positions
  - [0xDC00-0xDFFF] = 0x400 = 1024 positions
  - 1 048 576 car. complémentaires

• 0000000000000000xxxxxxxxxxxxyyyyyyyyyyyy

110110xxxxxxxxxxx

1<sup>er</sup> seizet d'indirection

110111yyyyyyyyyyy

2<sup>e</sup> seizet d'indirection

# 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)

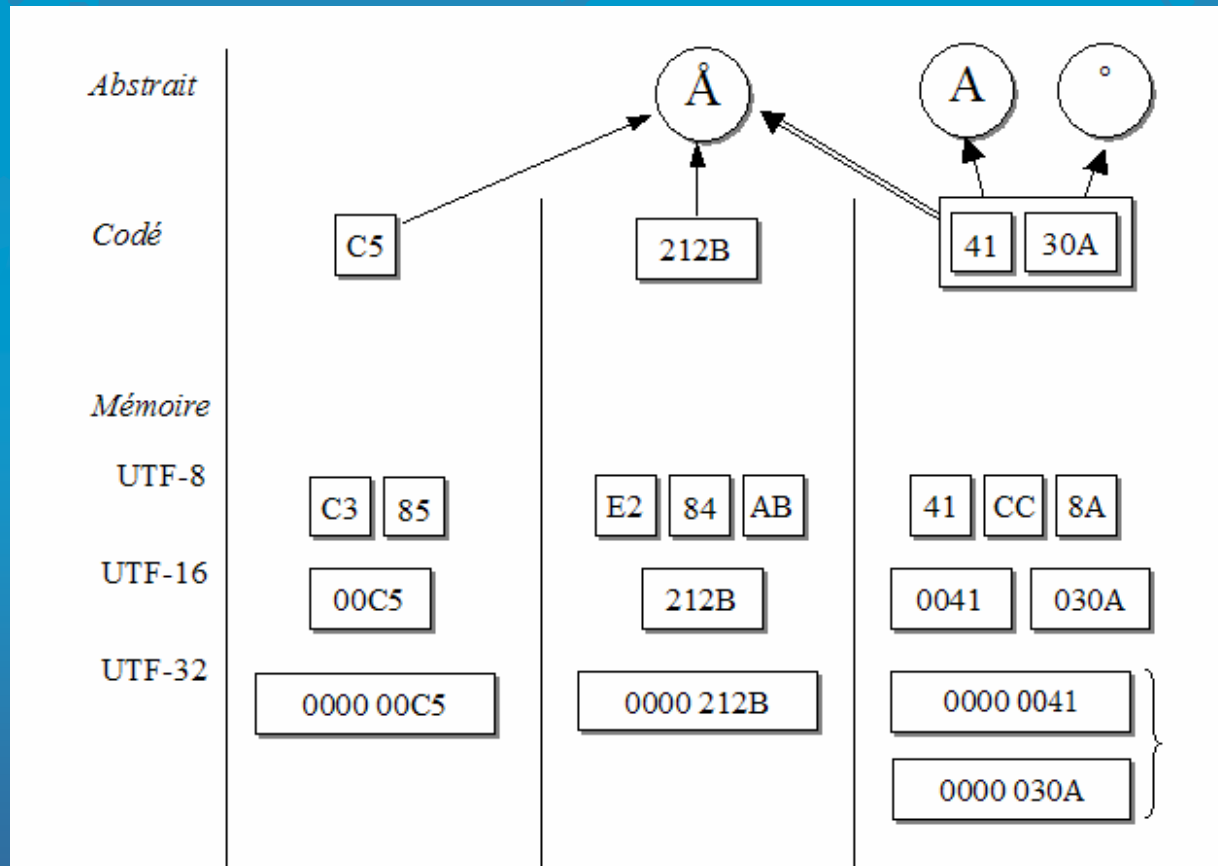
# UTF-8

Numéro de caractère	1 <sup>er</sup> octet	2 <sup>e</sup> octet	3 <sup>e</sup> octet	4 <sup>e</sup> octet
00000 00000000 0xxxxxxx	0xxxxxxx			
00000 00000yyy yyxxxxxx	110yyyyy	10xxxxxx		
00000 zzzzyyyy yyxxxxxxx	1110zzzz	10yyyyyy	10xxxxxx	
uuuuu zzzzyyyy yyxxxxxxx	11110uuu	10uuzzzz	10yyyyyy	10xxxxxx

# Exemples UTF-8

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

# Modèle de codage





# Sérialisation

- Le fait de transformer un groupe de bits, reçus en parallèle, en une succession de bits, transmis dès lors en série
- En pratique sérialisation en octets
- On va sérialiser les unités de codage (octets [UTF-8], seizets [UTF-16, UCS-2], « trente-deuzets » [UTF-32, UCS-4])
- Transfert
  - Petit-boutien (d'abord octets de plus petit poids)
  - Grand-boutien

# Sérialisation

<i>Sérialisé</i>			
UTF-8	C3 85	E2 84 AB	41 CC 8A
UTF-16BE	00 C5	21 2B	00 41 03 0A
UTF-16LE	C5 00	2B 21	41 00 0A 03
UTF-32BE	00 00 00 C5	00 00 21 2B	00 00 00 41 } 00 00 03 0A }
UTF-32LE	C5 00 00 00	2B 21 00 00	41 00 00 00 } 0A 03 00 00 }

# Surcodage

Le surcodage de transfert est une transformation réversible de données codées

- Éviter l'utilisation de certaines valeurs d'octets non compatibles avec les protocoles de transfert
- Appliquer différents algorithmes de compression de données :
  - SCSU (fenêtres)
  - BOCU (deltas, unicité de la compression, préserve ordre)

# 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...
  - ...

# Catégorie générale

A faint, light blue world map is visible in the background of the slide, centered behind the text.

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

# 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 (récurives) respectives sont identiques

# Classes combinatoires

- Différentes positions où s'attache les diacritiques
- Ces classes sont utilisées par l'algorithme de mise en ordre canonique défini par le standard Unicode
  - 0: Avec chasse, fendues, englobantes, antéposées
  - 1: Couvrantes et intérieures
  - 7: Nouktas
  - 208: Antéposées et jointes à gauche
  - 218: Souscrites à gauche
  - 224: Adscrites à gauche (d'un seul car de base)
  - 230: Suscrites
  - Etc...

# Babelmap

The screenshot displays the BabelMap application window. At the top, the title bar reads "U BabelMap" and the menu bar includes "Fichier", "Edition", "Rechercher", "Outils", and "Options ?". The main area is titled "Plan multilingue de base : Latin de base [0000..007F] (128 caractères)". It features a grid of characters organized by hexadecimal code points (0070 to 00E0) and columns (0 to F). The character "È" is highlighted in red in the grid at code point U+00C8. A large yellow callout box with a black border shows the character "È" in a large font. Below the grid, the text "U+00C8 : LETTRE MAJUSCULE LATINE E ACCENT GRAVE" is visible. The interface includes a search section with a dropdown menu set to "Latin de base", a search input field, and a "Rechercher" button. To the right, there is a field for "Aller au point de code" with the value "0000" and an "Aller" button. At the bottom, there is an "Edition" buffer containing the text "«Babelmap» : un utilitaire de visualisation de caractères Unicode". To the right of the buffer are buttons for "Effacer", "Copier", and "Enregistrer". The "Mode" section at the bottom left has radio buttons for "Caractère" (selected), "ACN (hexa)", "ACN (décimal)", and "NCU". The "Police" section at the bottom right shows "Arial Unicode MS" and a font size of "22".

# Propriétés

**Propriétés de caractère U+0020** [X]

Information générale

Nom de bloc : Latin de base

Nom du caractère : ESPACE

Nom Unicode 1.0 :

Commentaire ISO : Introduit dans version : 1

Propriétés fondamentales

Catégorie générale : Zs [Séparateur, espace]

Classe comb. canonique : 0 [Avec chasse, fendues, englobantes, antéposées et subjointes tibétaines]

Classe Bidi : WS [Blanc] Réflexion bidi ? Non

Décomposition et valeur numérique

Type de décomposition : Correspondance :

Type numérique : Valeur :

Notes [\*], synonymes [=] et renvois [x]

- \* le mot "espace" est féminin en typographie
- + autres espaces : 2000-200A
- x espace insécable - 00A0
- x espace sans chasse - 200B
- x gluon de mots - 2060
- x espace idéographique - 3000

Autres propriétés

Blanc  
Base de graphème

Variantes standardisées

Correspondance de casse

Casse minuscule :

Casse majuscule :

Casse de titre :

Codage

UTF-8 : 20

UTF-16 : 0020

UTF-32 : 00000020

Prononciation CJC

Copier OK

# Normalisation

- Afin de garantir une représentation unique de ce qui est considéré comme équivalent (canonique ou de compatibilité) car
  - Unicode définit parfois plusieurs codes qui correspondent à des entités peuvent être considérées comme identiques (variantes canonique) ou qui ne sont que des variantes visuelles d'un même caractère (variantes de compatibilité).

# Exemples

- On peut écrire de manière équivalente le mot été de la façon suivante :
  - É + t + é
  - E + é + t + é
  - E + é + t + e + é
  - É + t + e + é

# Exemples (suite)

Manières de coder LAM ALIF HAMZA EN CHEF ISOLÉ :

- ل (U+0644) + ا (U+0623)
- ل (U+0644) + ا (U+0627) + َ (U+0654)

Formes de compatibilités, déconseillées

- لا (U+FEF7)
- لا (U+FEFB) + َ (U+0654)

# Décomposition canonique

- Réversible
- N'entraîne aucune perte d'information.
- Elle peut donc être utilisée dans l'échange normalisé de textes.
- En effet, cette forme permet d'effectuer une comparaison binaire tout en conservant une équivalence canonique avec le texte non normalisé d'origine.

# Décomposition de compatibilité

- Perte d'information visuelle (pas exactement la même apparence)
- La *décomposition de compatibilité* permet d'effectuer une comparaison binaire tout en conservant cette fois-ci une équivalence de compatibilité avec le texte non normalisé d'origine.
- Peut s'avérer utile car elle permet d'éliminer des différences qui ne sont pas toujours pertinentes

# Noms des formes normalisées

	Sans composition canonique	Suivie d'une composition (C) canonique
Décomposition Canonique	D	C
Décomposition de compatibilité (K)	KD	KC

# Formes normalisées d'« affligé »

**Nom**

**Chaîne normalisée**

D

a + ffli + i + g + e + ó

C

a + ffli + i + g + é

KD

a + f + f + l + i + g + e + ó

KC

a + f + f + l + i + g + é

# Forme C et W3C

- Dans les recommandations du W3C
  - XML
  - XHTML
  - URL internationalisés
- Normalisation uniforme à la source
  - Source le plus à même (sait par exemple qu'on n'a affaire qu'à un sous-répertoire comme Latin-1)
  - Éviter de normaliser à toutes les étapes



*Merci*

*S*

# Ressources

- Introduction à Unicode, Patrick Andries  
<http://cooptel.qc.ca/~pandries/pdf/intro-Unicode.pdf>
- Fontes et codage, par Yannis Yaralambous, O'Reilly, 2004
- Unicode en français :
  - <http://cooptel.qc.ca/~pandries/pdf/Chapitre-4.pdf>
  - <http://cooptel.qc.ca/~pandries/pdf/Chapitre-6.pdf>
  - <http://pages.videotron.com/hapax/UCD-4.0.0.fr.html>
- Babelmap
  - [http://www.cooptel.qc.ca/~pandries/BabelMap\\_fr.html](http://www.cooptel.qc.ca/~pandries/BabelMap_fr.html)