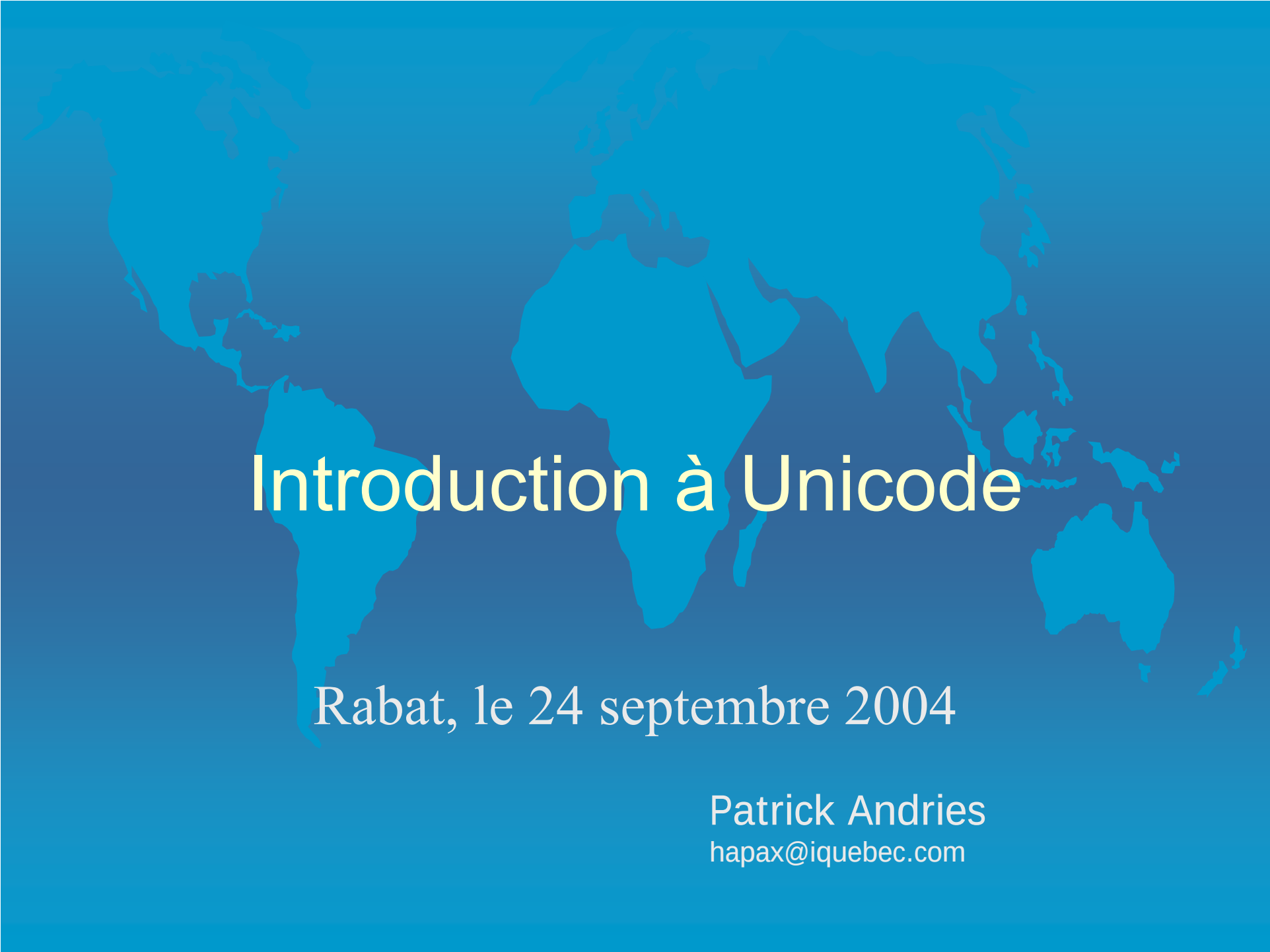


A stylized world map in a lighter shade of blue, centered on the Atlantic Ocean, serves as the background for the slide.

Introduction à Unicode

Rabat, le 24 septembre 2004

Patrick Andries
hapax@iquebec.com

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
- Ὑνιῳδοη
- য়ুনিকোড
- Unicode
- ynikod
- ا‏ن‏ك‏و‏د
- யூனிகோட்
- ယူနီကုဒ်
- ゆにこおど
- 유니코드
- 统一码
- 統一碼
- Γιούνικοντ
- Юникод
- ཡུ་ཤི་ཀོ་དྲ་
- यूनिकोड
- ဟွ်ဂုဒုလုဝ

Omniprésent

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

ISO/CEI 10646 et Unicode

Même répertoire :

- même caractères
- mêmes noms
- mêmes numéros

ISO/CEI 10646 et Unicode

- Unicode publié en anglais uniquement (bien qu'une traduction annotée en français d'Unicode 3.1 existe sur <http://pages.infinet.net/hapax>)
- Unicode définit plus de comportements (formes normalisées, coupures de lignes, algorithme bidi)
- Unicode fournit plus d'information sur les caractères et écritures (tri mais ISO 14651, casse, etc.)

ISO/CEI 10646 et Unicode

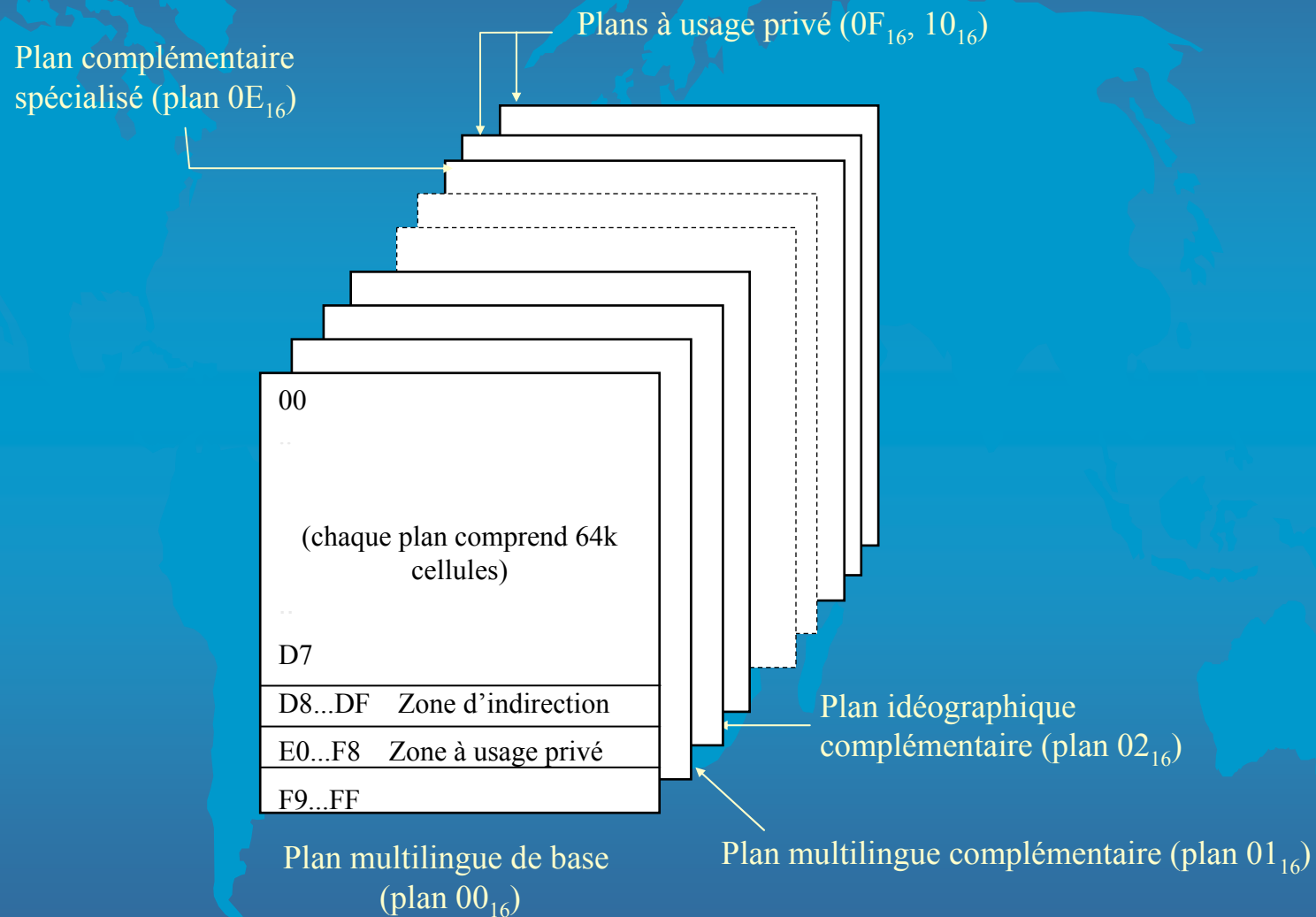
Membres :

- SC2
 - Organismes nationaux (membres votants)
 - SNIMA est membre 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

Plans de codage

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

Plans et zones de codage



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.

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 (ϖ , $\mathfrak{B}$, $\mathbb{B}$, $\mathfrak{M}$)

Caractères — abstraits, codés et glyphs

À	Ȧ	A	Ḑ	U+0041	lettre majuscule latine a
ff	ffi	ffi	ffi	U+0066	lettre minuscule latine f +
				U+0066	lettre minuscule latine f +
				U+0069	lettre minuscule latine i
ف	ف			U+0647	lettre arabe fa'

Caractères de base et diacritiques

Caractères

e ö ú ̲ ̇

Glyphe

→ è

Empilement de diacritiques

प + ि → पि

Signe vocalique indien antéposé (i bref)

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 天心 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

- *ل*, *ه* et *ح* sont tous des U+0062 *h*
- 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

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 : ̣ ̥ ̦ ̧

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>`

6. Ordre logique

En mémoire

M a r o c

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

À l’affichage

Maroc

ג ב נ ש

U+05E9 U+05E0 U+05D1 U+05D2

ש נ ב ג

- Utilise les propriétés de directionalité 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

8. Composition dynamique

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

9. Séquence équivalente

LJ + U

≡

L + J + U

A + ö + ◌

≡

A + ◌ + ö

A + ö + ó

≠

A + ó + ö

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

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

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



Merci

Ressources

- ISO et Unicode en français
 - <http://pages.infinit.net/hapax>
 - Tableaux de caractères
 - Textes explicatifs (plus de 300 pages)
- Liste de diffusion et forum de discussion
 - (Unicode Afrique)
 - <http://fr.groups.yahoo.com/group/Unicode-Afrique/>
 - (Unicode fr)
 - [nntp://fr.comp.normes.unicode](mailto:fr.comp.normes.unicode)