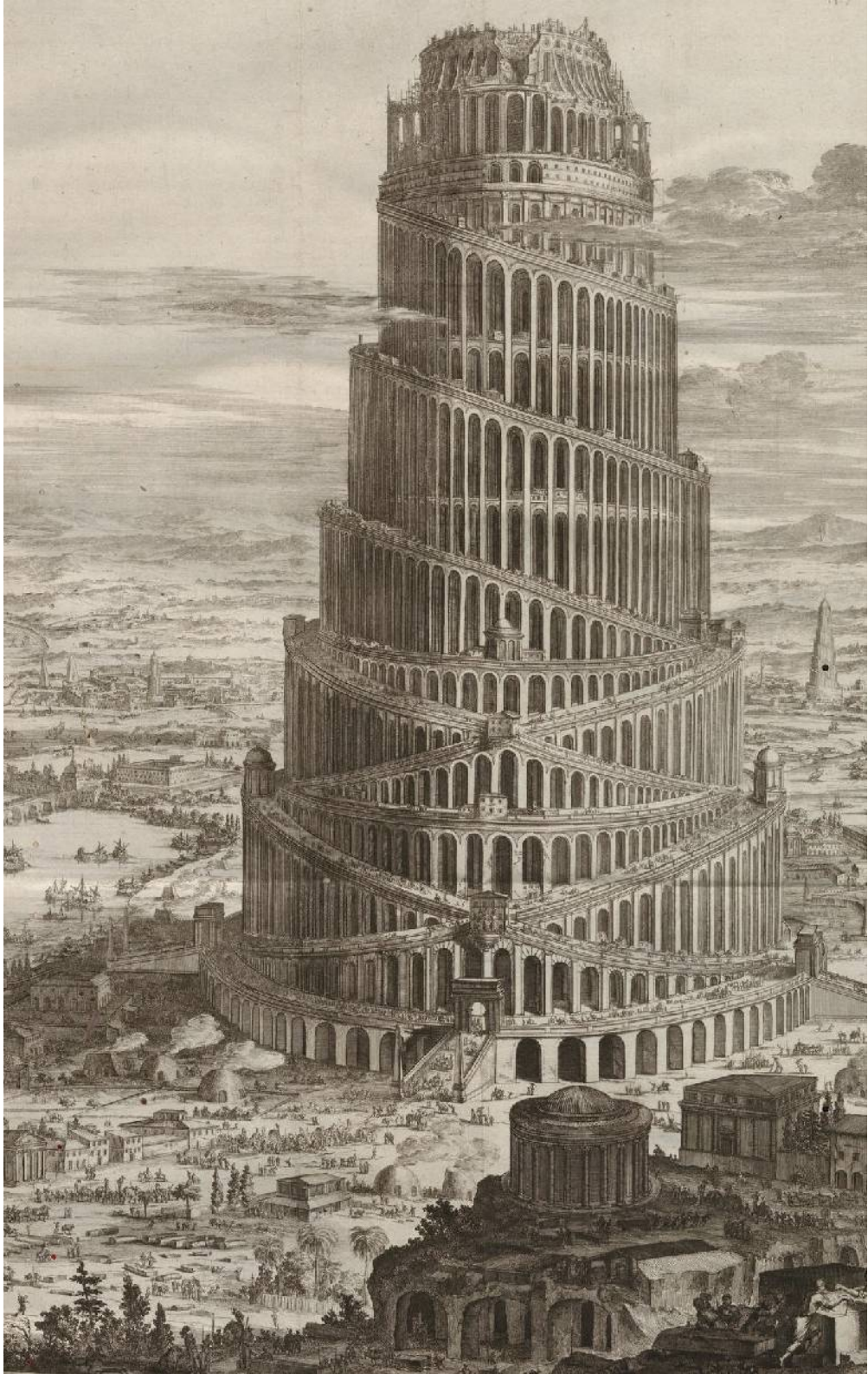
Unicode

Ateliers de l'Information
17 mars 2015
Guillaume Allègre



Plan de la présentation

Introduction – Unicode

- Unicode : un inventaire exhaustif
  - de tous les caractères
  - dans tous les systèmes d'écriture du monde

Introduction – Unicode

- Unicode : un inventaire exhaustif
  - de tous les caractères
  - dans tous les systèmes d'écriture du monde
- Qu'est-ce qu'un caractère ?

Introduction – Unicode

- Unicode : un inventaire exhaustif
  - de tous les caractères
  - dans tous les systèmes d'écriture du monde
- Qu'est-ce qu'un caractère ?
  - **U+0041 LATIN CAPITAL LETTER A**
  - 8 *glyphes* (en fait une infinité !)
  - un *codepoint* (U+0041)
  - un nom normalisé (en anglais)

Introduction - Propriétés

- **U+0041 LATIN CAPITAL LETTER A**
  - **http://unicode.org/cldr/utility/character.jsp?a=0041**

- Quelques propriétés
  - la casse : **U+0041 A** et **U+0061 a**
  - Bidi_Class :Left_To_Right
  - Block : Basic_Latin

Introduction - La forme et le fond

- La forme : **A A A**
- Le fond
  - **A**lphabet — U+0041 **LATIN** CAPITAL LETTER A
  - **Α**λφάβητο (en grec) — U+0391 **GREEK** CAPITAL LETTER ALPHA
  - **А**лфавит (en russe) — U+0410 **CYRILLIC** CAPITAL LETTER A

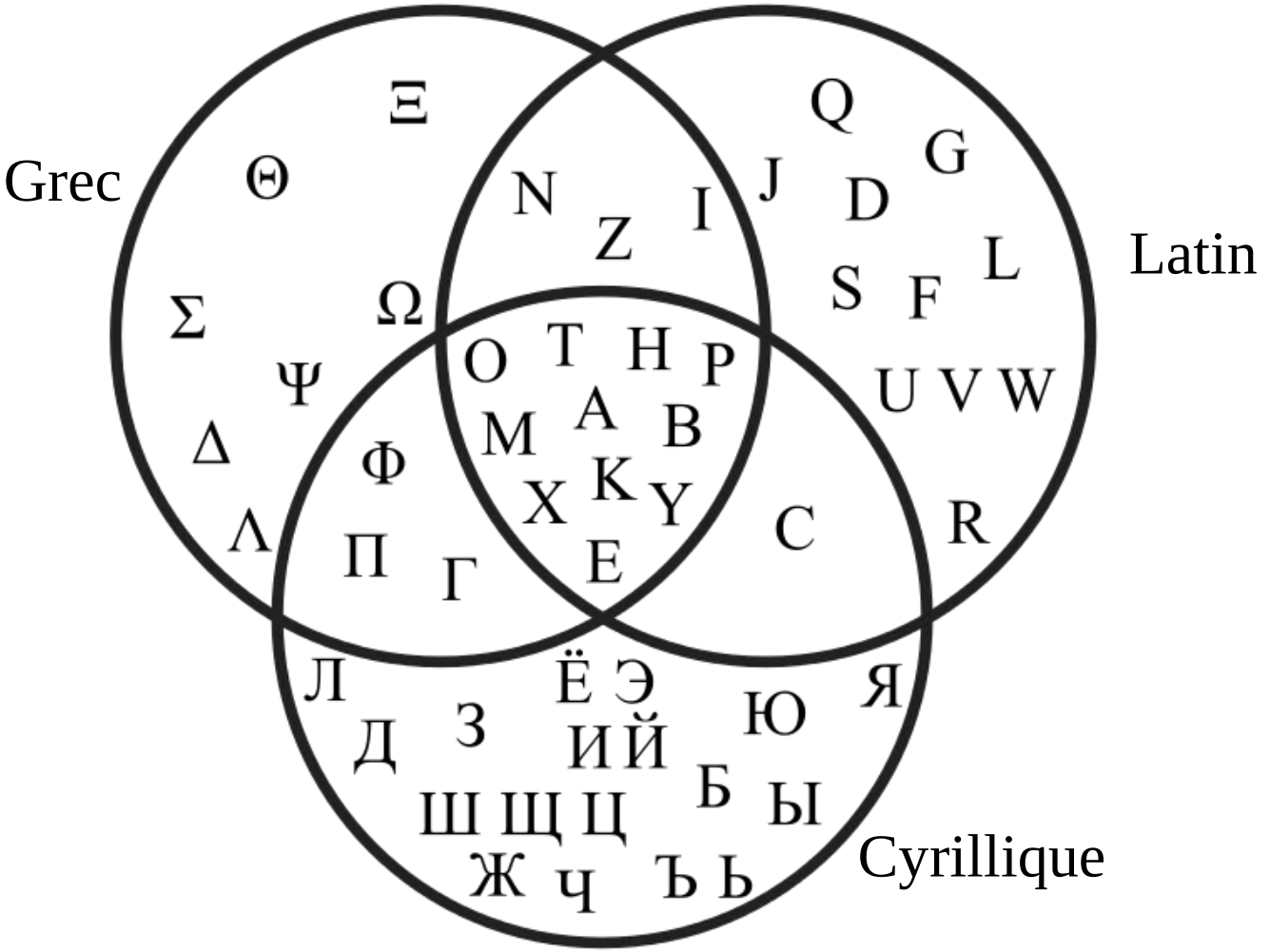
Introduction - La forme et le fond

- La forme : **A A A**
- Le fond
  - **A**lphabet — U+0041 **LATIN** CAPITAL LETTER A
  - **Α**λφάβητο (en grec) — U+0391 **GREEK** CAPITAL LETTER ALPHA
  - **А**лфавит (en russe) — U+0410 **CYRILLIC** CAPITAL LETTER A
- En bref
  - un seul glyphe (*dans la police de ce document*)
  - trois caractères différents

Introduction – Langues et écritures

- U+0041 LATIN CAPITAL LETTER A
  - **A**lphabet : français, anglais, allemand
  - **A**lfabeto : italien, espagnol, portugais
  - **A**akkoset : finnois
- U+0410 CYRILLIC CAPITAL LETTER A
  - **А**лфавит : russe
  - **А**бетка : ukrainien

Introduction – *une incise*



(les trois descendent de l'alphabet phénicien)

Introduction - Latin et latin étendu

- Un alphabet pour de nombreuses langues
- Conséquences
  - **Đ, ð, Ø, Þ, þ, ß, ı** + autres **lettres additionnelles,**
  - **é, à, ï, ñ, ô, ċ, İ** + autres lettres avec **signes diacritiques,**
  - ***Æ, æ, Œ, œ, ß*** + et autres **ligatures.**
  - **...**

En pratique : trois couches

- **(U) Jeu de caractères abstraits**
- (U) Encodages
  - représentation en mémoire des caractères
  - formats d'échange (fichier, réseau)
- Affichage
  - polices de caractères
  - logiciels d'affichage (bibliothèques)

Historique 1

- Codes ASCII (1963 -)
  - codé sur 7 bits (un caractère = un mot mémoire)
  - 128 caractères (33 de contrôle + 95 affichables)
  - adapté à l'alphabet anglais (latin, sans diacritique)

ASCII Code Chart

	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	CAN	EM	SUB	ESC	FS	GS	RS	US
2		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
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

Historique 2

- « ASCII *étendu* » : plusieurs ***jeux de caractères***
  - codage sur 8 bits : + 128 valeurs (128-255)
  - *codepages* IBM 1980-
    437 Standard,
    850 Multilingual Latin-1...
  - ISO-8859-* (1986 -)

⚠ pas de multilinguisme,
pas de vraie typographie...

Historique 3 – Unicode

- 1991 : **Unicode 1.0** 65536 caractères (possibles)
  - 1992 : Unicode 1.0.1 + 20902 CJK Unified Ideographs
  - 1993 : Unicode 1.1
  - 1993 : ISO 10646-1:1993 francisation
- 1996 : **Unicode 2.0** 1 112 064 caractères. (+1)
- 1999 : Unicode 3.0 (+ 3.1, 3.2)
- 2003 : Unicode 4.0 (+ 4.1)
- 2006 : Unicode 5.0 (+ 5.1, 5.2)
- 2010 : Unicode 6.0 (+ 6.1, 6.2, 6.3)
- 2014 : Unicode 7.0
- ...

Unicode – encodages

I  Unicode

* U+FFFD REPLACEMENT CHARACTER – bloc *Specials* (U+FFF0..U+FFFF)

Unicode - encodages

- Une erreur historique
  - Unicode 1.0 : 65536 caractères, 16 bits
  - *Wide chars* de largeur fixe, ex. UCS-2
  - Unicode 2.0 : ***plans*** supplémentaires, +1M caractères
- Encodages principaux
  - **UTF-8 standard web, compatible ASCII**
    ex. L'Ã©tÃ© est lÃ : erreur d'encodage
  - UTF-16
  - UTF-32

Inventaire des systèmes d'écriture

- ISO 15924
  - une norme secondaire à Unicode (2004 -)
  - 171 systèmes inventoriés à ce jour
  - 22 révisions (dernière 2014-11-15) ...
- Pour chaque écriture
  - un code alphabétique sur 4 lettres, ex. “Latn”
  - un numéro sur 3 chiffres, ex. 215
  - un nom anglais, ex. “Latin”
  - un nom français, ex. “Latin”

ISO 15924 – Les dix séries

codet	écritures	exemple
0xx	écritures hiéroglyphiques et cunéiformes	090 Maya
1xx	écritures alphabétiques de droite à gauche	160 Arab
2xx	écritures alphabétiques de gauche à droite	290 Teng (pas dans Unicode)
3xx	écritures alphasyllabiques	315 Deva
4xx	écritures syllabiques	411 Kana
5xx	écritures idéographiques ou symboliques	500 Hani
6xx	écritures non déchiffrées	620 Roro
7xx	sténographie et autres notations	760 Dupl
8xx	série pas encore utilisée	
9xx	codets à usage privé, codets spéciaux	997 Zmth

L'écriture arabe – 1

- La plus proche des écritures « exotiques » ?
- Cinq différences par rapport au latin
  - écriture de droite à gauche
  - pas de voyelles
  - pas de différence de casse (majuscules/minuscules)
  - 4 formes positionnelles
  - ligature entre les caractères

L'écriture arabe – 2 – similarités

- De nombreuses langues l'utilisent
  - arabe, farsi, sindhi (Arab/Deva), kurde (Arab/Cyrl/Latn) ...
  - turc jusqu'en 1928
- Nombreuses adaptations nécessaires
- Une écriture liturgique !

L'écriture arabe – un exemple

Coran bilingue arabe et persan (noir, incliné), diacritiques (rouge), 1250-1350
https://commons.wikimedia.org/wiki/File:Qur%27an_manuscript_Surat_al-Nisa%27._%281%29.tif

L'écriture arabe – direction

- S'écrit de droite à gauche

يولد جميع الناس أحرارًا متساوين في الكرامة والحقوق

- Difficultés
  - Textes bilingues

    Le texte ci-dessus utilise le mot الناس pour traduire « *les hommes* »
  - Nombres et dates

التاريخ 14 تموز 1789م

پس از کودتای **۲۸** مرداد سال **۱۳۳۲**، محمدرضا شاه به تثبیت قدرت خود پرداخت

Après le coup d'Etat du 28 *mordad* 1332, le Shah a consolidé son pouvoir. *(WP-fa)*

L'écriture arabe – direction

- Gestion de la ***directionnalité*** en Unicode
  - propriété d'un caractère
    ex. BIDI Right-to-Left Arabic [AL]
  - deux caractères de contrôle pour la forcer :
    U+200E LEFT-TO-RIGHT MARK
    U+200F RIGHT-TO-LEFT MARK
  - une norme **très** complexe pour gérer l'affichage

L'écriture arabe – pas de voyelles

- Un ***abjad***
  - les lettres sont toutes des consonnes
  - les voyelles sont représentées par des diacritiques
- Nombreuses adaptations aux langues autres que l'arabe
  - lettres ajoutées
  - diacritiques (points...) suscrits, souscrits... pour noter les voyelles

L'écriture arabe – formes

- Pas de casse (majuscule / minuscule)
- Quatre formes de position
  - Forme isolée, ex. (nūn) ن U+FEE5
  - Forme initiale, نـ U+FEE7
  - Forme médiale, ـنـ U+FEE8
  - Forme finale, ـن U+FEE6
  - Forme « indifférenciée » (standard) U+0646

L'écriture arabe – formes

- Pas de casse (majuscule / minuscule)
- Quatre formes de position
  - Forme isolée, ex. (nūn) ﻥ U+FEE5
  - Forme initiale, ﻧ U+FEE7
  - Forme médiale, ﻨ U+FEE8
  - Forme finale, ﻦ U+FEE6
  - Forme « indifférenciée » (standard) U+0646
- En latin ?
  - *... le deſir de vous amuſer, voilà les ſeuls motifs ...*
  - U+017F LATIN SMALL LETTER LONG S

L'écriture arabe – ligature automatique

- Exemple

  يولد جميع الناس أحرارًا متساوين في الكرامة والحقوق

- Unicode
  - ـ U+0640 ARABIC TATWEEL (ligature)
- Autres technologies
  - bibliothèque d'affichage (modification de texte)
- En latin ?
  - *... le deſir de vous divertir eſtoit les ſeuls motifs ...*
  - U+FB05 LATIN SMALL LIGATURE LONG S T

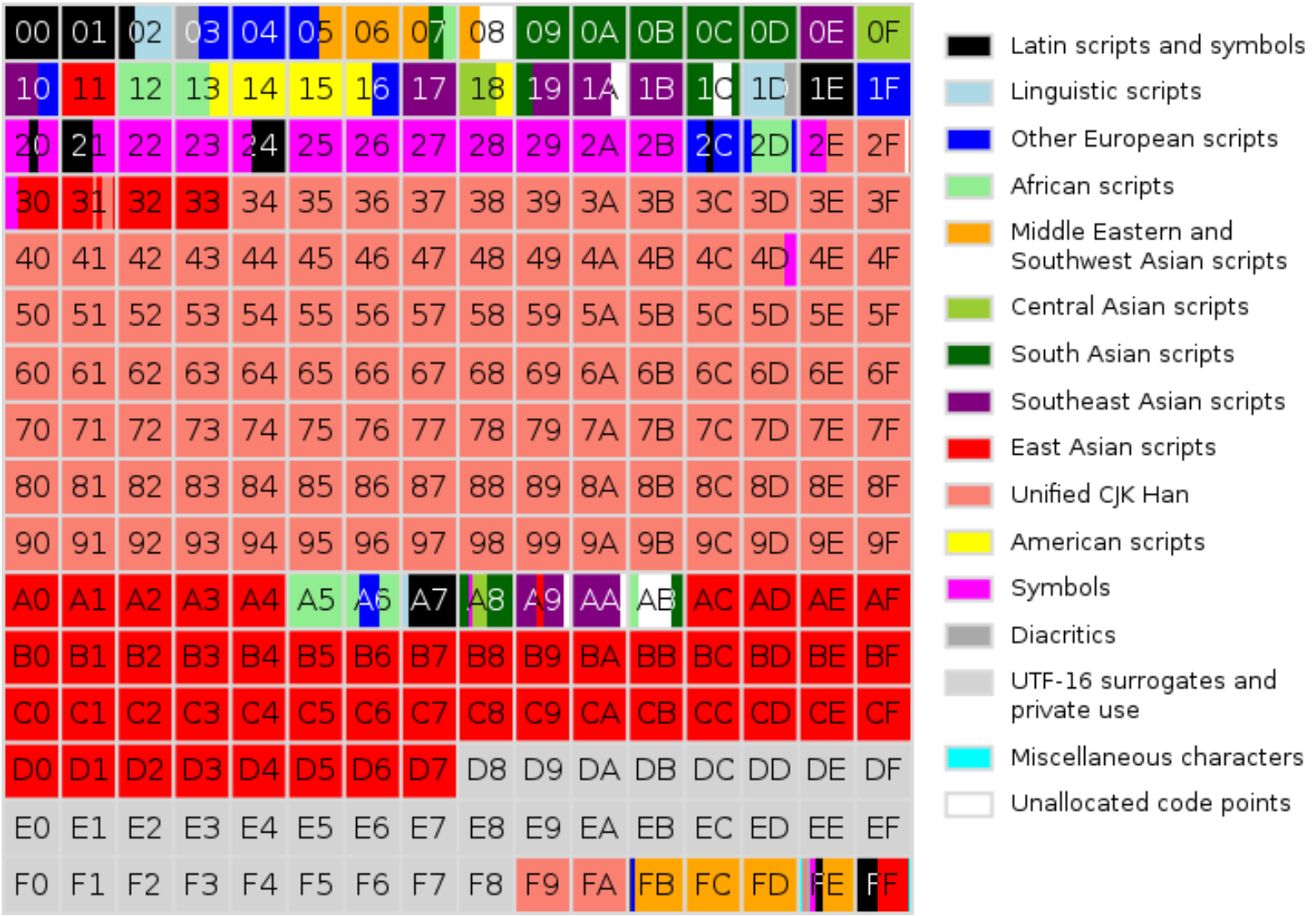
Unicode – organisation 1

- 17 ***plans***
  - Numérotés de 0 à 10 (en hexadécimal)
  - De 65536 caractères chacun (0000 – FFFF)
  - 1 114 112 *codepoints* possibles
- En particulier
  - 0 (BMP) Basic Multilingual Plane
  - 1 (SMP) Supplementary Multilingual Plane
  - 2 (SIP) Supplementary Ideographic Plane
  - …
  - E (SSP) Supplementary Special-purpose Plane

Unicode – organisation 2

- Chaque plan est divisé en ***blocs***
  - **unité** d'allocation (administration Unicode)
  - intervalle de *codepoints*, nnn0 – nnnF (multiple de 16)
  - un nom unique
- Exemples
  - C0 Controls and Basic Latin (0000–007F)
  - Arabic (0600–06FF)
  - Hangul Syllables (AC00–D7AF)

Organisation du plan 0 (BMP)



https://commons.wikimedia.org/wiki/File%3ARoadmap_to_Unicode_BMP.svg
Par Saric [domaine public], via Wikimedia Commons

Unicode : traitement du texte

- Collation
  - tri (ordre « alphabétique »)
    dépend de la langue
  - comparaison, équivalence de chaînes
    « l'été est là » == « l'ete est la »
  - Forme ***normale*** et ***composition*** : **é**
    U+00E9 LATIN SMALL LETTER E WITH ACUTE
    U+0065 LATIN SMALL LETTER E **+** **U+0301** COMBINING ACUTE ACCENT
- Normalisation des traitements automatiques
  - capitalisation, ex.
    Diyarbakır → DIYARBAKIR → diyarbakir (général)
    Diyarbakır → DİYARBAKIR → diyarbakır (turc)

Unicode et son contexte technologique

- Polices de caractères
  - déclarent une liste de caractères couverts
- Affichage du texte
  - composition, ligatures, fontes
- Méthodes de saisie
  - périphériques : clavier
  - claviers virtuels, phonétique...

Unicode – Art

- Smileys ASCII
  - occidentaux :-) ;-p
  - orientaux (^_^) (o_O) m(._.)m
- Smileys Unicode
  - simples : ♥‿♥
  - sophistiqués : (͡° ͜ʖ ͡°)

Symboles

- De nombreux symboles normalisés
-    
-     
-     
-     
- ...

Emojis

- Une généralisation des smileys (émoticônes)
  - 
  - 
- Apparus très tôt au Japon
  - 1998 NTT DoCoMo (opérateur mobile)
  -
- Normalisés tardivement dans Unicode
  - Smartphones : Google (Android) et Apple

Emojis – 2

- Sur-représentation de la culture japonaise
  -  **U+1F30D**-F EARTH GLOBE...
  - Mont Fuji  , Tour de Tokyo  , Carte du Japon  **U+1F5FB**-
  - Poupées  , château  , ogre  , goblin  « japonais »

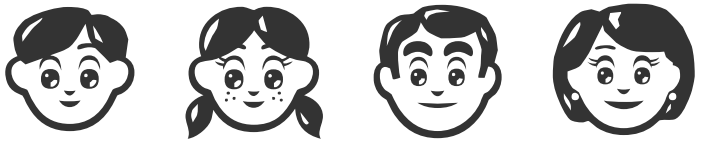
Unicode pour tous

- Exemple 
  - U+1F46B MAN AND WOMAN HOLDING HANDS
  - U+1F46C TWO MEN HOLDING HANDS
  - U+1F46D TWO WOMEN HOLDING HANDS

Unicode pour tous (ou presque)

- Exemple 
  - U+1F46B MAN AND WOMAN HOLDING HANDS
  - U+1F46C TWO MEN HOLDING HANDS
  - U+1F46D TWO WOMEN HOLDING HANDS
- mais 
  - U+1F46A FAMILY.

Unicode pour tous (ou presque)

- Exemple 
  - U+1F46B MAN AND WOMAN HOLDING HANDS
  - U+1F46C TWO MEN HOLDING HANDS
  - U+1F46D TWO WOMEN HOLDING HANDS
- mais 
  - U+1F46A FAMILY.
- Au passage : 
  - U+1F466 - 1F469 BOY, GIRL, MAN, WOMAN

Conclusion ?

- Une face technique
  - Unicode reference charts http://unicode.org/charts/
- Une face récréative
  - Shapecatcher http://shapecatcher.com/

Crédits – licence CC-By-SA 3.0