Numérisation des écritures

Michel Bottin

michel.bottin@free.fr

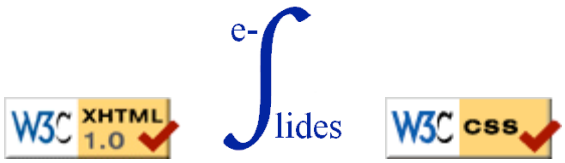

Table des matières

30. Signets

Définitions

Écriture :
au sens strict : ensemble de signes représentant des éléments de langue. C'est une **logographie**, système graphique de notation du langage.
Généralement, une écriture donnée sert à noter plusieurs langues :
- l'écriture arabe pour l'arabe, le persan, l'ourdou, le ouïghour, etc.
- l'écriture chinoise pour le chinois, partiellement le coréen et le japonais
- l'écriture cyrillique pour le russe, l'ukrainien, le bulgare, etc.
- l'écriture latine pour le français, l'anglais, l'allemand, le swahili, le vietnamien, le quechua, etc.

Système d'écriture :
un ensemble de signes d'une écriture associé à des règles de combinaison et de disposition graphique propres à une langue donnée.

Apparition des écritures

Dès le Paléolithique :

- des incisions ou marques
- des signes (tectiformes, claviformes, paillottes, etc.)

Écritures proprement dites. Cinq foyers (indépendants ?) :

- Mésopotamie : Uruk (act. Warka) dans le pays de Sumer [-3300]
- Égypte [-3100]
- Harappa : civilisation de l'Indus [-2600]
- Chine : sous la dynastie des 商 (Shāng) [-1400]
- Amérique centrale : Olmèques, Maya et Zapotèques [-400], Nahuatl (Aztèques) [1200]

Carte

Type d'écritures

Si l'on écarte les "écritures" mythographiques ou pictographiques qui sont plutôt des notations de discours, la double articulation du langage (1: morphèmes, 2:phonèmes) correspond à deux grands types d'écritures :

- **morphémographie**, un signe graphique dénote une *unité linguistique signifiante*. Ex. : l'écriture chinoise
- **phonographie**, un signe graphique dénote une *unité linguistique non-signifiante*
  - **syllabe**, dénotation d'un groupe de sons prononçables en une seule émission de voix. Ex. : les kanas (syllabaires japonais)
  - **phonème**, dénotation des sons :
    - uniquement les consonnes : alphabets consonnantiques ou abjads. Ex. : les alphabets arabe, hébraïque ou syriaque
    - les consonnes et les voyelles (voire les tons). Ex. : les alphabets cyrilliques, grecs,

latins

Disposition

Linéarité des écritures :

- les écritures sont toujours composées d'éléments de base concaténés. Cette disposition linéaire découle du caractère unidimensionnel de la chaîne parlée.
- **direction** : la concaténation peut se faire suivant toutes les directions : de gauche à droite, de droite à gauche, de haut en bas, de bas en haut, etc.
- **segmentation** : comme une ligne d'écriture ne peut pas avoir une longueur infinie, elle est le plus souvent *segmentée*. Généralement les différents segments sont placés les uns au dessous des autres dans le cas de directions horizontales, les uns à gauche ou à droite des autres pour dans le cas des directions verticales. C'est la *page*.
  Il a existé des cas d'écritures — grec archaïque — où la direction alternait suivant les segments βουστροφηδόν (boustrophédon) avec fréquemment le retournement des lettres d'une ligne à l'autre.
- **reliure** : les pages sont généralement attachées par le côté proche du début de la première ligne pour constituer des *livres*. Mais il existe d'autres modes d'organisation : en rouleaux, en accordéon, etc.

Structure de surface et structure profonde

Les éléments de la chaîne linéaire qui paraissent en surface d'un niveau donné, peuvent parfois se décomposer en éléments de niveau inférieur.
Comparons ces deux séries :

	hiragana			**devanāgarī**		
	/a/	/i/	/u/	/a/	/i/	/u/
/p/	ぱ	ぴ	ぷ	प	पि	पु
/t/	た	ち	つ	त	ति	तु
/k/	か	き	く	क	कि	कु

hiragana : aucun élément commun ni verticalement ni horizontalement
devanāgarī : l'élément de la colonne /a/ se retrouve dans les deux autres cellules de la même ligne.
Un autre élément est présent dans chacune des deux dernières colonnes.

Comme il en est de même pour les autres combinaisons consonne(s)-voyelles on peut affirmer que :

- l'hiragana est un véritable syllabaire
- la devanāgarī a une **structure de surface** syllabique et une **structure** profonde alphabétique

Typologie détaillée des écritures

On distinguera donc des écritures :

- **morphémographiques** (improprement appelées idéographiques ou logographiques)
- **phonographiques**
  - **syllabiques**

        - **syllabiques** : structure de surface syllabique indécomposable
        - **alpha-syllabiques** : structure de surface syllabique et structure profonde phonémique (abugidas ou Hangul Jamo)
    - **phonémiques**
        - **défectives** : uniquement la composante consonnantique (abjads)
        - **complètes** : les composantes consonnantique et vocalique et éventuellement tonale (alphabets)

Principe généraux du codage

La numérisation des écritures suppose toujours les étapes suivantes :

- déterminer le répertoire des signes, c'est à dire les caractères abstraits (graphèmes ?) par opposition aux glyphes
- ordonner ce répertoire en utilisant un ordre traditionnel ou créé pour la circonstance
- affecter à chaque élément de ce répertoire une valeur numérique respectant éventuellement l'ordre choisi

Des questions techniques vont se poser suivant le nombre de caractères et la correspondance entre les classes abstraites que sont les caractères et les glyphes concrets

Caractères (graphèmes) versus Glyphes

- **glyphes** : variantes graphiques non systématiquement pertinentes (usages pratiques ou esthétiques)
  a, a, a, a, a, a , ,
- **graphème** : caractère abstrait, distinctif à l'intérieur du système d'écriture.
  Pour Unicode les glyphes ci-dessus sont subsumés dans le caractère nommé LATIN SMALL LETTER A

Mais des glyphes peuvent être identiques et pourtant correspondre à des caractères distincts car appartenant à des systèmes d'écritures distincts :
A : LATIN CAPITAL LETTER A (U+0041)
А : CYRILLIC CAPITAL LETTER A (U+0410)
Α : GREEK CAPITAL LETTER ALPHA (U+0391)

La distinction est parfois difficile à faire. Cf. l'unification des caractères Han.

Codification des écritures (1/2)

Le développement de l'informatique avait conduit à une véritable *balkanisation* : pas de correspondance bi-univoque entre codes et caractères

Ex. le code 232 (E8 en hexadécimal) correspond à :

- **è** en ISO 8859-1 (Latin 1)
- **č** en ISO 8859-2 (Latin 2)
- ш en ISO 8859-5 (Cyrillique)

- و en ISO 8859-6 (Arabe)
- θ en ISO 8859-7 (Grecque)
- etc., etc.

D'où la nécessité d'attribuer un code numérique unique à chaque graphème (caractère).

Codification des écritures (2/2)

En revanche, nous verrons qu'en Unicode, les caractères précédents auront tous des valeurs distinctes :

- **è** : conserve la valeur 232 (E8 en hexadécimal)
- **č** :a la valeur 269 (10D en hexadécimal)
- ш :a la valeur 1096 (448 en hexadécimal)
- و :a la valeur 1608 (648 en hexadécimal)
- θ :a la valeur 952 (3B8 en hexadécimal)

Les caractères distincts ont des valeurs distintes. Aucune confusion n'est possible.

Histoire d'Unicode / ISO 10646

Enfin Unicode vint...

- **1986-1987** : début de l'unification Han chez Xerox et du jeu de caractères universel chez Apple (projet Apple File Exchange).
- **1987-12** : Première mention d'**Unicode** (pour unique, universel, uniforme) dans un document
- **1989** : DP 10646 publié indépendamment. Premier brouillon d'Unicode publié par Apple, Claris, Metaphor et Sun. Présentation à IBM et Microsoft.
- **1990** : publication d'Unicode 1.0
- **1991** : Unicode et ISO/IEC 10646 s'accordent pour fusionner. Création du Consortium Unicode sous le nom "Unicode, Inc".
- **1993** : Unicode 1.1 révisé pour coïncider avec l'ISO 10646-1:1993
- **1996** : publication de l'Unicode 2.0
- **1998** : publication de l'Unicode 2.1
- **1999** : publication de l'Unicode 3.0
- **2003** : publication de l'Unicode 4.0 (actuellement 4.0.1)

Pour en savoir plus...

Principes généraux d'Unicode

Le standard Unicode a pour principes :

- l'**universalité** : Unicode doit comporter tous les caractères utilisés dans l'échange de textes
- **unicité** : chaque valeur de groupe de 16 bits représente un caractère et un seul
- l'**uniformité** un code de caractère de longueur fixe permet d'effectuer de manière efficace tris, recherches, affichage et édition de textes

Et aussi :

- l'**efficacité** : un texte simple composé de caractèress de longueur fixe est facile à analyser. Nul besoin de maintenir des états, d'analyser des séquences d'échappement ou de faire des recherches progressives ou régressives
- le **codage des caractères et non pas des glyphes** : nombre fini vs indéterminé

Caractéristiques d'Unicode

- **caractères de 16 bits** : UCS-2, UTF-16, UTF-8, ...
- **encodage complet** : > 96000 caractères dans la version 4.0 (et ce n'est pas fini...)
- **caractères, pas les glyphes** : nombre fini vs indéterminé
- **sémantique** : numérique, espacement, combinaison, directionnalité, ...
- **texte brut** : caractères, directionnalité, langue
- **ordre logique** : ordre de la structure profonde et pas de la structure de surface
- **unification** : de 130000 "idéogrammes" à 27786 codes !
- **composition dynamique** : existence de caractères combinatoires. Ex. : e + ´ → é
- **séquences équivalentes** : tout caractère précomposé est décomposable. Ex. : ã ≡ a + ˜
- **convertibilité** : avec les standards préexistants sans perte d'information

Unicode n'est pas une **grammatologie** scientifique mais un effort pragmatique et réaliste réussi !

Représentation des caractères

Cinq niveaux :

- **répertoire abstrait de caractères (ACR)** : description pure
- **jeu de caractères codés (CCS)** : valeurs numériques associés
- **forme naturelle de caractères (CEF)** : représentation informatique pure
- **schéma d'encodage de caractères (CES)** : mécanismes de sérialisation, UTF-xx, CESU-8, SCSU, BOCU, etc.
- **syntaxe d'encodage de transfert (TES)** : "imprimable balisé", "base 64", ...

Structure des standards Unicode/ISO 10646 : les plans

(128 groupes de 256)(16) plans de 256 rangées de 256 caractères :

- 0 - **BMP** : Basic Multilingual Plane = *Plan de base multilingue*
- 1 - **SMP** : Supplementary Multilingual Plane = *Plan supplémentaire multilingue*
- 2 - **SIP** : Supplementary Ideographic Plane = *Plan supplémentaire idéographique*
- 14 - **SSP** : Supplementary Special-purpose Plane = *Plan supplémentaire spécial*
- 15 : *Réservé pour usage privé*
- 16 : *Réservé pour usage privé*

Plan 0 : BMP, le plan de base multilingue

De U+000000 à U+00FFFF, les blocs d'allocation :

- U+000000-U+001FFF : **Ecritures générales**
- U+002000-U+002E7F : **Symboles**
- U+002E80-U+0033FF : **Phonogrammes et Symboles CJK**
- U+003400-U+009FFF : **Idéogrammes CJK**
- U+00A000-U+00A4CF : **Syllabes et clé Yi**
- U+00AC00-U+00D7AF : **Syllabes Hangul**
- U+00D800-U+00DBFF : **Substitutions**
- U+00E000-U+00F8FF : **Privés**
- U+00F900-U+00FFFF : **Compatibilité et spéciaux**

Plan de base multilingue (détail)

Plan 1 : SMP, le plan supplémentaire multilingue

De U+010000 à U+01FFFF, les blocs d'allocation :

- U+010300-U+01032F : **Italiques antiques**
- U+010330-U+01034F : **Gotique**
- U+010400-U+01044F : **Deseret**
- U+01D000-U+01D0FF : **Symboles musicaux byzantins**
- U+01D100-U+01D1FF : **Symboles musicaux occidentaux**
- U+01D400-U+01D7FF : **Symboles mathématiques alphanumériques**

Plan supplémentaire multilingue (détail)

Plan 2 : SIP, le plan supplémentaire idéographique

De U+020000 à U+02FFFF, les blocs d'allocation :

- U+020000-U+02A6DF : **Idéogrammes CJK - Extension B**
- U+02F800-U+02FA1F : **Idéogrammes de compatibilité CJK supplémentaires**

Plan 14 : SSP, le plan supplémentaire spécial

De U+0E0000 à U+0EFFFF, les blocs d'allocation :

- U+0E0000-U+0E007F : **Étiquettes** (marques de langue)

Plans 15 et 16 : les deux plans d'usage privé

- U+0F0000-U+0FFFFD : **Zone d'usage privé supplémentaire - A**
- U+100000-U+10FFFD : **Zone d'usage privé supplémentaire - B**

Les encodages

Du code de caractère à sa représentation :

- **UCS-2** : fixe 16 bits (2^{16} = 65.536 caractères)
- **UCS-4** : fixe 31 bits (2^{31} = 2.147.483.648 caractères)
- **UTF-7** : variable 7 bits [obsolète]
- **UTF-8** : variable 8 bits (un caractère occupe de 1 à n octets). Peut représenter tous les caractères de l'UCS-4
- **UTF-16** : semi-fixe 16 bits / 32 bits (2^{20} = 1.048.576 caractères ; un caractère du BMP occupe 2 octets, les caractères des 15 plans suivants utilisent une paire haute et basse de caractères de substitution)
- **UTF-32** : fixe 32 bits (identique à l'UCS-4 pour le domaine de l'UTF-16)

- UCS : Universal Caracter Set = Jeu de caractères universel
- UTF : UCS Transformation Format = Format de transformation de l'UCS

Mise en œuvre

La codification résout la représentation des écriures à l'intérieur des fichiers.
Mais demeurent plusieurs problèmes :

- comment saisir ces caractères ? question des claviers ou plus généralement des méthodes d'entrée
- comment effectuer des ordonnancements ? question des algorithmes de tri
- comment les afficher ou imprimer ? question des polices de caractères

La solution de ces questions relève des systèmes d'exploitation et/ou des applications.

Systèmes d'exploitation et Unicode

Introduction:

- Java UCS-2 natif, support des substituts depuis Java 1.4
- Linux : OpenI18N (Open Internationalization Initiative) [http://www.openi18n.org/]
  Παν語 (Pango) http://www.pango.org/
- MacOS
  - 9 : avec les kits de langues
  - X : support des substituts
- Windows
  - 95 et 98 : BMP seulement
  - 2000 et XP : support des substituts

Claviers et méthodes d'entrée

- Les systèmes d'exploitation récents possèdent tous des dispositifs pour les "principales" langues.
- *A noter* : MacOS X a défini une définition de clavier en XML compilable a la première utilisation.

- Java a défini une interface pour les éditeurs de méthode entrée (IME) des applications développées avec ce langage.
- La communauté des **Logiciels libres** se devrait de développer une définition de clavier et/ou de méthode d'entrée compilable sur toutes les plateformes

Polices de caractères

- Polices généralistes :
  - Arial Unicode MS (51.180 glyphes Unicode 2.0 en version 0.86, 23Mo !)
  - Bitstream Cyberbit (29.934 glyphes en version 2.0β)
  - Code2000 (34.654 glyphes en version 1.12)
  - Code2001 (caractères du plan supplémentaire multilingue)
  - TITUS Cyberbit Basic (9560 glyphes en version 2.1)
- Polices spécifiques à une écriture et/ou langue :
  - polices latines, grecques, cyrilliques, arabes, hébraïques
  - polices CJKV de style spécifique (chinois simplifié, chinois traditionnel, coréen, japonais)
  - polices indiennes
  - pour des écritures moins répandues

Un projet : **SmartFont**, sous licence libre de polices, selon le standard SVG (Scalable Vector Graphic) du W3C, indépendant des plateformes.

Conclusion

Utilisez toujours Unicode UTF-8 !

- universel
- compatibilité maximale avec les applications existantes
- indépendance vis à vis des plateformes matérielles

Des questions ?

FIN

Merci de votre attention !

Liens

- Carte [apparitions.html]
- Pour en savoir plus... [http://www.unicode.org/unicode/history/]
- Plan de base multilingue (détail) [p000-bmp.html]
- Plan supplémentaire multilingue (détail) [p001-smp.html]

Signets

- Écritures du monde [http://www.culture.fr/edm/]
- Unicode Home Page [http://www.unicode.org/]
- World Wide Web Consortium [http://www.w3.org/]
- Unicode 3.1 et ISO 10646 en français [http://iquebec.ifrance.com/hapax/]
- ISO 15924 - Codes d'écritures [http://www.evertype.com/standards/iso15924/]
- Roadmap to BMP, SMP, SIP ans SSP [http://std.dkuug.dk/JTC1/SC2/WG2/docs/n2461.pdf]
- Character Encoding Model [http://www.unicode.org/unicode/reports/tr17/]
- FAQ - UTF and BOM [http://www.unicode.org/unicode/faq/utf_bom.html]
- Authoring Techniques for XHTML & HTML International 1.0 [http://www.w3.org/TR/2003/WD-i18n-html-tech-20031009/]
- Alan Wood's Unicode Resources [http://www.alanwood.net/unicode/]
- Les cinq livres [http://www.unicode.org/Public/TEXT/FIVEBOOKS]
- Convertisseur Unicode [http://site.voila.fr/fllcjvg/unicode.htm]

Aide à la navigation

Vue	Clavier	Barre de navigation
Suivant	'N' or ' ' or Enter	>
Précédent	'P'	<
Title slide	'T'	>
Premier	'F'	|>
Dernier	'L'	<|
Contenu	'C'	C
Liens	'K'	K
Signets	'B'	B
Aide	'H' or '?'	?

Barre de navigation dans le site (Mozilla 1.x)