Table des matières

Avant-propos XVII

Première partie – Introduction

Chapitre 1 – Concepts de base et terminologie 3
1.1 Pourquoi Unicode ? 3
1.2 Absence de codage universel 5
1.3 Langue et écriture 6
1.3.1 Notation, écriture, transcription et translittération 7
1.3.2 Qu'est-ce qu'un caractère ? 8
1.3.3 Graphème, caractère et glyphe 9
1.3.4 Terminologie de l'apparence des caractères 11
1.3.5 Classification des caractères 11
1.3.6 Caractères romains, latins, italiques et gothiques 12
1.3.7 Écriture CJC 13
1.4 Unicode, en quelques mots 14
1.4.1 Ce qu'Unicode est… 14
1.4.2 Ce qu'Unicode n'est pas 14
1.5 Apprivoiser les polices Unicode 16
1.5.1 Afficher des caractères Unicode 16
1.5.2 S'assurer que sa police est une police Unicode 17
1.5.3 Où trouver des polices multi-écritures supplémentaires ? 18
1.5.4 Absence de glyphe pour un caractère 20
1.5.5 Incorporation des polices 21
1.6 Saisir des caractères Unicode 22
1.6.1 Claviers 25

1.6.2 Méthodes d'entrée extrême-orientales 28
1.6.3 Tableau de caractères 29
1.7 Internationalisation et localisation 32

Chapitre 2 – Répertoires et jeux de caractères codés 35
2.1 Terminologie 35
2.1.1 Répertoire de caractères 36
2.1.2 Jeu de caractères codés, code et codage 36
2.2 L'ASCII 37
2.2.1 ISO 646 et « variantes nationales de l'ASCII » 39
2.2.2 « ASCII 8 bits » 39
2.3 Les codes ISO/CEI 8859 40
2.3.1 ISO/CEI 8859-1 ou ISO Latin-1 41
2.3.2 ISO/CEI 8859-15 ou ISO Latin-9 43
2.4 Windows Latin 1 44
2.5 EBCDIC 46
2.6 KOI8-R 47
2.7 ISO/CEI 2022 47
2.8 ISO/CEI 10646 et Unicode 49
2.9 GB 18030, l'« Unicode chinois » 51

Deuxième partie – L'essentiel d'Unicode

Chapitre 3 – Structure d'Unicode 55
3.1 Principes directeurs 55
3.2 Caractères normalisés 56
3.2.1 Plan multilingue de base 57
3.2.2 Plans complémentaires 61
3.2.3 Nombre de caractères normalisés 61
3.2.4 Unicode 5.1 62
3.3 Caractères 64
3.3.1 Caractère abstrait et caractère codé 64
3.3.2 Caractères combinatoires et diacritiques 65
3.3.3 Suite de caractères de base et diacritiques 66
3.3.4 Caractères combinatoires multiples 66
3.4 Principes de conception d'Unicode 67
3.4.1 Universalité 68

3.4.2 Efficacité 69
3.4.3 Caractères et non glyphes 70
3.4.4 Sémantique 70
3.4.5 Texte brut 71
3.4.6 Ordre logique 71
3.4.7 Unification 72
3.4.8 Composition dynamique 72
3.4.9 Stabilité 74
3.4.10 Convertibilité 75

Chapitre 4 – Modèle de codage, propriétés des caractères et tri 77
4.1 Modèle de codage des caractères 77
4.1.1 Répertoire de caractères abstraits 80
4.1.2 Jeu de caractères codés 80
4.1.3 Mot, octet, seizet, codet 80
4.1.4 Forme en mémoire des caractères 81
4.1.5 Mécanisme de sérialisation de caractères 84
4.2 Propriétés des caractères 86
4.2.1 Voir les propriétés grâce à BabelMap 86
4.2.2 Catégorie générale 88
4.2.3 Casse 91
4.2.4 Classe combinatoire canonique 94
4.2.5 Directionalité 95
4.2.6 Réflexion bidi 96
4.2.7 Classes de coupure de lignes 96
4.2.8 Coupure de lignes et segmentation de texte 99
4.2.9 Cohérence des propriétés 100
4.3 Formes normalisées 101
4.3.1 La concaténation n'est pas fermée 103
4.3.2 Formes de normalisation et base de données 103
4.3.3 Stabilité des formes normalisées 103
4.3.4 Invariance des caractères latins de base 104
4.3.5 Compositions exclues 104
4.4 Le tri et le repérage 105
4.4.1 Tri et comparaison binaires 106
4.4.2 Tri et comparaison lexicographiques 106
4.4.3 Ça ne peut quand même pas être si compliqué ? 106
4.4.4 La solution – un tri à quatre niveaux 107

4.4.5 Les éléments de tri et les clés de tri 109
4.4.6 La table DUCET 110
4.4.7 Expansions et contractions 111
4.4.8 Mise en œuvre du tri Unicode 112
4.4.9 Personnalisation 113
4.5 Conformité 113
4.6 Le standard Unicode : mode d'emploi 115
4.6.1 Liste des noms de caractère 115
4.6.2 Images dans les tableaux et dans les listes de caractères 116
4.6.3 Renvois 116
4.6.4 Renseignements sur les langues 117
4.6.5 Décompositions 117

Troisième partie – Caractères remarquables

Chapitre 5 – Lettres et signes diacritiques 123
5.1 Latin étendu et API 123
5.2 Lettres modificatives 124
5.3 Clones à chasse des diacritiques 125
5.4 Écriture grecque 125
5.5 Signes diacritiques 127
5.5.1 Diacritiques généraux 128
5.5.2 Supplément de diacritiques 129
5.5.3 Diacritiques destinés aux symboles 129
5.5.4 Demi-signes diacritiques 130

Chapitre 6 – Ponctuation 131
6.1 Ponctuation Latin-1 131
6.1.1 Guillemet anglais 132
6.1.2 Croisillon 132
6.1.3 Perluète 132
6.1.4 Apostrophe 132
6.1.5 Astérisque et obèle 133
6.1.6 Trait d'union-signe moins 134
6.1.7 Arrobe 134
6.1.8 Clones de diacritique 134
6.1.9 Paragraphe et pied-de-mouche 135
6.1.10 Symbole degré et ordinal masculin 136

6.1.11 Point médian 137
6.1.12 Trait d'union conditionnel 138
6.1.13 Ponctuation appariée 140
6.2 Guillemets 140
6.2.1 Usages européens 140
6.2.2 Usage extrême-oriental 142
6.3 Espaces 142
6.3.1 Espace mot et espace insécable 143
6.3.2 Les différents caractères d'espacement 144
6.3.3 Ajustement de l'espacement 146
6.3.4 Espaces fines en français 146
6.3.5 Espace sans chasse 149
6.3.6 Disposition des espaces 149
6.4 Autres signes typographiques 151
6.4.1 Points de suspension et points de conduite 151
6.4.2 Traits d'union et tirets 152
6.4.3 Puces, barre de fraction, ponctuation doublée 154
6.4.4 Ponctuation archaïque 154
6.5 Caractères de coupure de lignes 155

Chapitre 7 – Symboles et notations 157
7.1 Symboles de type lettre 157
7.2 Symboles monétaires 159
7.3 Mathématique 159
7.4 Musique 161

Chapitre 8 – Caractères techniques spéciaux 163
8.1 Caractères de commande 163
8.1.1 Commandes C0 et suppression 163
8.1.2 Commandes C1 164
8.2 Gluon et diacritique invisible bloquant 165
8.2.1 Gluon de mots (U+2060) 165
8.2.2 Diacritique invisible bloquant (U+034F) 165
8.3 Caractères spéciaux 166
8.3.1 Délimiteurs d'annotation interlinéaire 167
8.3.2 Non-caractères 168
8.4 Positions non attribuées 168

8.5 Caractères déconseillés et désuets 169
8.6 Zones à usage privé 169
8.7 Indicateur d'ordre des octets 170
8.8 Étiquettes linguistiques 171

Quatrième partie – Applications et techniques liées à Unicode

Chapitre 9 – Préciser la langue, l'écriture et le pays 177
9.1 ISO 639 – indicatifs de langue 178
9.2 ISO 3166 – indicatifs de pays 180
9.3 M.49 – Indicatifs de pays et de régions 181
9.4 ISO 15924 – indicatifs d'écriture 181
9.5 RFC 4646 – Étiquettes de langue 182
9.7 BCP 47 184

Chapitre 10 – Unicode et les protocoles Internet 187
10.1 De l'utilité des métadonnées 187
10.2 Les premiers protocoles Internet 188
10.3 Type de médias Internet/type MIME 189
10.3.1 Visualiser les entêtes 189
10.3.2 Les types de média ou types de contenu 191
10.3.3 L'information sur le codage de caractères (« charset ») 192
10.3.4 Les entêtes relatifs au surcodage de transfert 193
10.3.5 Le surcodage des entêtes 197
10.3.6 Recettes de dépannage 198
10.4 Codage de caractères sur le Web 200
10.4.1 Entêtes HTTP 200
10.4.2 Préciser le codage de caractères dans HTTP 201
10.4.3 Vérifier les entêtes HTTP 202
10.4.4 Quel codage utiliser pour mes pages web ? 204
10.4.5 Balise meta 204
10.4.6 Préciser le codage dans XHTML et XML 205
10.4.7 Conflits des définitions de codage 206
10.4.8 Configuration de serveurs web 207
10.5 HTTP internationalisé 210
10.5.1 La négociation de langue 210
10.5.2 Les entêtes reliés aux caractères 219

10.6 Adresses internationalisées 220
10.6.1 Noms de domaine et DNS 220
10.6.2 Internationaliser les URI 222
10.6.3 Noms de domaine internationalisés (NDI) 224
10.6.4 Menaces informatiques : hameçonnage et parodie 226
10.6.5 Caractères non ASCII dans les chemins des IRI 229
10.7 La locale POSIX 232

Chapitre 11 – Unicode et (X)HTML, XML, CSS 239
11.1 Préciser le codage en (X)HTML 239
11.1.1 HTML 240
11.1.2 XHTML – le prologue XML 240
11.1.3 Préciser le codage en CSS 244
11.2 Préciser la langue 245
11.2.1 HTML et XML 245
11.2.2 Passages dans une autre langue 245
11.2.3 Documents bilingues 246
11.2.4 La langue dans l'entête HTTP ou l'attribut lang ? 247
11.3 Préciser la directionalité 248
11.4 Stylage sensible à la langue 249
11.4.1 Les sélecteurs CSS 249
11.4.2 Utilisation des sélecteurs de langue 250
11.5 Schémas XML internationalisés 251
11.5.1 Texte dans les attributs XML 251
11.5.2 Les éléments qui ne contiennent qu'une chaîne 252
11.5.3 Prévoir l'attribut `xml:lang` 252
11.5.4 Prévoir un élément de type `span` 252
11.5.5 Ne pas créer d'éléments de présentation 252
11.5.6 Prévoir `xml:id` *sur tous les éléments traduisibles* 253
11.6 Notation des caractères 253
11.6.1 Appels d'entités HTML : souvent de peu d'utilité 255
11.6.2 Appels de caractère : à n'utiliser que rarement 256
11.6.3 Quand les appels de caractère et d'entité sont utiles 256
11.6.4 Entités en XHTML 258
11.7 Caractère ou balisage ? 258
11.7.1 Unicode contient trop de caractères 258
11.7.2 Caractères de commande en HTML et XHTML 259

11.7.3 Autres caractères permis et interdits en XML 260
11.7.4 Caractères incompatibles avec le balisage 261
11.7.5 Caractères de compatibilité Unicode 263
11.8 Réglage de l'algorithme bidi 263
11.8.1 Bref rappel de l'algorithme bidi 264
11.8.2 Désactiver l'algorithme 268
11.8.3 Cas problématiques – les neutres mal placés 269
11.8.4 Éditer du texte bidi balisé 270
11.8.5 Emboîtement des passages bidi 271
11.8.6 Caractères de commande bidi et balisage 273
11.8.7 Les feuilles de styles CSS et le bidi 273
11.9 Formulaire « universel » 275
11.9.1 Créer un formulaire 275
11.9.2 Accept-charset sur la balise form 278
11.9.3 Inclusion de caractères étrangers au `charset` 279
11.9.4 Solution : n'envoyer et n'accepter que de l'UTF-8 279

Chapitre 12 – Internationalisation des logiciels 283
12.1 Internationaliser ? 283
12.1.1 PNB par langue 283
12.1.2 L'adaptation de logiciels à la pièce 284
12.1.3 Internationalisation 284
12.1.4 Quelques a priori culturels 285
12.1.5 Localisation ou adaptation culturelle 286
12.2 Les langages de programmation 287
12.2.1 Les caractères dans C et C++ 287
12.2.2 Les caractères dans Java 289
12.2.3 Les caractères dans C# et la plateforme .NET 291
12.2.4 Bibliothèques d'internationalisation, le cas ICU 292
12.2.5 Propriétés de caractères en Java/ICU 293
12.2.6 Unicode dans les expressions régulières 295
12.2.7 UnicodeSet en ICU 296
12.2.8 Normalisation 297
12.2.9 Comment supprimer les accents d'un texte ? 298
12.2.10 Créer des noms de domaine internationalisés 299
12.2.11 Les transformations ICU 299
12.3 Profil culturel ou locale 302
12.3.1 Concept et nécessité 302

12.3.2 Définir une Locale 303
12.3.3 Définir une Locale avec ICU 303
12.3.4 Locale sur .NET 304
12.3.5 Locale implicite 305
12.3.6 Changer la casse d'une chaîne de caractères 305
12.3.7 Correspondance de casse non localisée 306
12.4 Isoler les données culturelles 307
12.4.1 Pourquoi extraire ? 307
12.4.2 Les « ResourceBundle » 307
12.4.3 Définition de ResourceBundle 308
12.4.4 Accès aux ResourceBundle 309
12.4.5 Découverte et recherche des ResourceBundle 309
12.4.6 Où sont stockées les ressources ? 310
12.5 Formater les messages 312
12.5.1 Formats prédéfinis de date et heure 312
12.5.2 Formater la date et l'heure à l'aide de motifs 313
12.5.3 Formater et analyser chiffres et montants 316
12.5.4 Variabilité de l'ordre des mots dans les langues 318
12.5.5 Messages variables et internationalisés 319
12.5.6 Gestion de l'accord en nombre 322
12.6 Comparaison et tri 325
12.6.1 Tri en Java ou ICU pour Java 325
12.6.2 Comparer des chaînes en ignorant les accents 327
12.6.3 Personnaliser le tri 327
12.6.4 Ignorer la ponctuation 330
12.6.5 Tri des tableaux de données à plusieurs champs 331
12.6.6 Améliorer la performance 333
12.7 Frontières de texte 334
12.7.1 La classe BreakIterator 334
12.7.2 Exemples : détecter les frontières de phrase et de mot 336
12.7.3 Personnaliser la détection de frontières 338
12.8 CLDR 339
12.9 Les exceptions et l'i18n 340
12.10 Conversion de données 342
12.10.1 Entrées/sorties 342
12.10.2 Écrire de l'Unicode avec des OutputStream 344

12.10.3 String.getBytes() et l'analyseur XML 345
12.10.4 Maîtrise de la conversion grâce à java.nio 346
12.11 L'interface utilisateur 347
12.11.1 Foisonnement du texte 347
12.11.2 Neutralité culturelle 348
12.11.3 Internationalisation des images et des icônes 349
12.12 À ne pas internationaliser 350

Chapitre 13 – Unicode et les polices 355
13.1 Caractères et variantes de glyphes 355
13.2 Sélecteurs de variante 356
13.3 Impact sur le choix de police 356
13.3.1 Police de repli 357
13.3.2 Police de substitution 357
13.3.3 Police liée 358
13.3.4 CSS et ses « polices liées » 358
13.3.5 Utiliser un équivalent canonique 359
13.4 Ligatures 359
13.4.1 Liant et antiliant 360
13.4.2 Liaison cursive 362
13.4.3 Liant, antiliant et les écritures brahmies 363
13.4.4 Filtrage des liants et antiliants 364
13.4.5 Liant et antiliant dans les polices 364
13.5 Pas d'expédients ASCII, de l'Unicode ! 365
13.6 Passage des caractères aux glyphes 366
13.6.1 Des caractères et non des glyphes 366
13.6.2 Fonctionnement d'un moteur de rendu 367
13.6.3 Les polices de glyphes 371
13.7 Processus de rendu 378
13.8 Un moteur de rendu : Uniscribe 380
13.9 Adobe et Uniscribe 382
13.10 Fonctionnalités et règles OpenType 382
13.11 Intégrer le tout 385

Bibliographie 389

Index 391