Les jeux de caractères en XML et (X)HTML – la norme Unicode

Andrei Popescu-Belis
TIM / ETI, Université de Genève

Cours n°5

Problème : utilisation de différents alphabets dans les documents informatiques

- Systèmes d'écriture
  - **alphabétiques** : dizaines de signes
    - souvent phonétiques (origine commune ? Ougarit ?)
    - ex. : alphabet latin (et dérivés), grec, arabe, hébreu
  - **syllabiques** : centaine de signes
    - ex. : syllabaires japonais ou inuit

  - **idéographiques** : dizaines de milliers de signes
    - chronologiquement les plus anciens, « signes ≈ mots »
    - ex. : idéogrammes chinois/japonais

- Comment les utiliser dans les documents ?
- Comment être sûr qu'ils sont affichés correctement ?

Comment les ordinateurs affichent-ils les caractères ?

- Représentation machine : bits (*binary digits*), 0 ou 1
- Traduction des codes machine en lettres
  - réalisée par une interface graphique vers un écran
  - correspondance entre les octets (série de 8 bits) et les caractères (abstraits)
  - ajout d'informations sur la police (forme concrète)
- Exemple : 01000001 (→ 65) → A → A

Définitions

- Jeu de caractères (*character set*)
  - correspondance abstraite entre des caractères (définis conceptuellement) et des nombres (codes) dans un tableau
  - valeurs des nombres : 0–127 ou 0–255 ou 0–65535
- Encodage ou codage (*encoding*)
  - représentation concrète, dans l'ordinateur, des codes des caractères sous la forme d'une suite de bits 0/1
- Police de caractères (*font*)
  - jeu de caractères + forme de chaque caractère (glyphe)

Remarques

- Du point de vue du contenu des documents, la notion la plus importante est le jeu de caractères
- En principe, un document (page web, email, document XML) ne peut utiliser qu'un jeu de caractères, signalé au début du document, mais plusieurs polices de caractères

Trois jeux de caractères importants

- US-ASCII : 128 caractères
  - alphabet latin sans diacritiques, chiffres, etc.
- Famille « ISO latin » : 256 caractères possibles
  - alphabet latin avec caractères accentués, etc.
  - notation : ISO-8859-*X* ou parfois ISO-LATIN-*X*
  - **variantes** selon les groupes de langues
    - Europe de l'Ouest : ISO-8859-1, Centrale : ISO-8859-2
    - mais aussi grec 7, arabe 6, hébreu 8, cyrillique 5, thaï 11, etc.
- Unicode : environ 100'000 caractères
  - regroupe tous les alphabets en même temps
    - (encodages possibles : UTF-8, UCS-2, UCS-16, etc.)

La norme Unicode

- Un jeu de caractères normalisé
  - un caractère est représenté par deux octets = 65'536 possibilités
    - (maintenant quatre, si nécessaire, pour dépasser ce nombre)
- Dans un document utilisant le jeu de caractères Unicode, on peut utiliser simultanément
  - la plupart des caractères alphabétiques existants ou ayant existé
  - de nombreux caractères symboliques (p.ex. mathématiques)
  - des dizaines de milliers d'idéogrammes CJK
  - … et il reste encore des codes non affectés
- Consultation : http://www.unicode.org
  - surtout : http://www.unicode.org/charts
  - application Windows « Table de caractères » (*charmap.exe*)
  - fonction Word « insérer caractères spéciaux »

7

Historique de la norme Unicode

- Consortium Unicode + ISO (Org. int. de normalisation)

➔ Standard Unicode ou norme ISO/IEC 10646

- Version 1.0 (1991) : 28'302 caractères (sur 2 octets)
- Version 4.0 (2003) : 96'382 caractères (sur 4 octets)
- Version 5.0 (2006) : 98'884 caractères graphiques

1991	Unicode 1.0	= CD2 de ISO 10646-1
1993	Unicode 1.1	= ISO 10646-1
1996	Unicode 2.0	= ISO 10646-1 + amendements
2000	Unicode 3.0	= ISO 10646-1 1e édition
2002	Unicode 3.2	= ISO 10646-2 2e édition
2003	Unicode 4.0	= ISO 10646:2003 (3e version)
2006	Unicode 5.0	sera publié officiellement en 2007

8

Utilisation des jeux de caractères dans les traitements de texte

- La plupart des logiciels « propriétaires » n'indiquent pas de façon transparente le jeu de caractères utilisé
  - l'utilisateur peut seulement choisir des polices
  - l'utilisateur peut insérer des symboles
    - à partir d'une police
    - en utilisant le clavier avec les langues installées (Windows)
- Explication
  - le format de stockage du document est de toute façon indéchiffrable pour l'utilisateur, le logiciel prend en charge la gestion des caractères
- Problème
  - comment le document sera-t-il vu par d'autres utilisateurs ?
  - réponse : cela dépend des polices installées, des versions du logiciel, etc.

9

Solution : déclaration d'encodage en XML

- XML = format texte + balises
- Le texte utilise un certain jeu de caractères, enregistré avec un certain encodage = l'encodage réel du fichier dans la machine
- L'entête XML déclare explicitement cet encodage
  - `<?xml encoding="UTF-8"?>` ou
  - `<?xml encoding="ISO-8859-1"?>` , etc.
  - UTF-8 est l'encodage par défaut (si aucune valeur n'est précisée)
- Il faut être sûr que la déclaration de l'encodage dans le fichier et le format réel du fichier soient bien les mêmes
  - dans ce cas, le texte du fichier sera affiché correctement partout

10

Comment enregistrer un fichier XML avec un encodage précisé ?

- XML Spy
  - choisir File > Encoding > …, puis enregistrer le fichier
- MS Word
  - enregistrer au format texte brut
  - puis fenêtre (riche) pour le choix explicite de l'encodage
- Notepad ou Wordpad
  - les options d'enregistrement au format texte brut permettent de choisir l'encodage du fichier courant
    - Notepad ☺ : ANSI [=1 octet], Unicode [=UCS-2/UTF-16], Unicode big endian [=UTF-16BE], et UTF-8
    - Wordpad ☹ : texte [ANSI], texte MS-DOS, texte Unicode [UCS-2] et… RTF (pas un format texte)

11

Pour vérifier l'encodage réel d'un document au format XML ou texte

- Essayer de l'ouvrir avec un navigateur web
  - le menu Affichage > Encodage des caractères indique l'encodage par défaut « deviné » par le navigateur
  - le menu Affichage > Code source permet de voir s'il y a une déclaration de l'encodage (ou jeu de caractères)
  - on peut modifier l'encodage utilisé par le navigateur avec le menu Affichage > Encodage jusqu'à ce que l'aspect du document soit convenable (tous les caractères affichés correctement)
- Ouvrir le document avec MS Word en mode « texte brut »
  - le logiciel présente une fenêtre de dialogue assez riche
  - permet le choix de l'encodage supposé
  - montre ses conséquences sur l'aspect du document

12

Insertion de caractères spéciaux dans les documents XML

- Utilisation directe des caractères disponibles au clavier
- Copier/coller depuis d'autres documents
  - problème d'affichage dans XML Spy
- Utilisation des cinq entités prédéfinies :

```
&amp;   &    &lt;  <     &gt;  >
&quot;  "    &apos;  '
```

- Utilisation d'entités numériques de la forme ɛ ou ɛ (pour ε) avec les codes de caractères Unicode
- Déclaration de nouvelles entités dans une DTD
  - exemple : `<!ENTITY gamma "&#x3B3;">`
  - ➔ ce sera surtout utile avec la DTD de (X)HTML

Déclaration de l'encodage pour les documents XHTML et HTML

- XHTML : on peut le faire comme en XML
- HTML et XHTML : déclaration dans l'entête

```
<head>
  <meta http-equiv="Content-Type"
        content="text/html; charset=jeu"/>
  … … …
</head>
```

- Encodages (jeux de caractères) possibles
  - le plus souvent : UTF-8 ou ISO-8859-1 ou US-ASCII
  - liste : http://www.iana.org/assignments/character-sets
  - l'encodage déclaré doit être le même que l'encodage réel

Insertion de caractères spéciaux dans les fichiers (X)HTML

- Utiliser l'une des méthodes vues pour les documents XML
  - utilisation des caractères disponibles au clavier
  - copier/coller depuis d'autres documents
  - utiliser les entités numériques (p.ex. ɛ ou ɛ)
- Utiliser les entités de la DTD publique de (X)HTML, p.ex.

```
· &eacute;   &otilde;   &uuml;   &copy;
  &amp;      &lt;      &oelig;   &gamma;
```

  - des entités existent pour les caractères ISO-LATIN-1, d'autres caractères spéciaux et ponctuations, symboles, etc.
    - http://www.w3.org/TR/REC-html40/sgml/entities.html
    - http://www.w3.org/TR/xhtml1/DTD/xhtml-lat1.ent, xhtml-special.ent, xhtml-symbol.ent (fragments de DTD)
  - indiquées aussi dans les « Entry Helpers » de XML Spy

Conclusion

- En principe, les programmes se chargent de la gestion de l'encodage pour vous, sauf :
  - quand les bons encodages ne sont pas installés
  - quand le programme ne trouve pas le bon encodage
  - quand vous voulez partager vos fichiers dans un format autre que le format propriétaire initial
  - quand le programmeur, c'est vous !
    - i.e., en XML, (X)HTML, etc.
- Comprendre ce qui se passe est toujours utile...

Exercices (1/2)

- Reprenez votre page XHTML du cours 5 sous XML Spy
  - attribuez-lui la DTD officielle de XHTML 1.0 (cf. cours 5)
- Enregistrez la page au format UTF-8 (avec XML Spy) et déclarez aussi ce changement dans l'entête de la page
- Insérez un maximum de texte utilisant des caractères spéciaux dans plusieurs alphabets
  - p.ex. liste de noms originaux de villes que vous aimeriez visiter
  - comment insérer ces caractères spéciaux ? (rappel)
    - entités (X)HTML indiquées par XML Spy en bas à droite
    - entités (X)HTML numériques (p.ex. é pour é)
    - utilisation du clavier avec sélection de la langue (Windows)
    - copier/coller depuis le web
      (problèmes possibles d'affichage sous XML Spy, tester avec Firefox)

Exercices (2/2)

- Vérifiez que la page s'affiche correctement avec Firefox et IE, sur votre poste et sur un poste voisin
- « Nettoyez » votre page pour lui donner un aspect présentable
- Validez la page à http://validator.w3.org en indiquant son URL
  - si elle est valide, faites comme indiqué (sur le site) pour afficher le logo « *Valid*! » et le lien permettant la validation ultérieure par n'importe quel visiteur
- Devoir à rendre: indiquez sur moodle l'URL du résultat http://home.etu.unige.ch/~votrelogin/xxxxxx.yyyy ou http://home.etu.unige.ch/~votrelogin (voir note qui suit)

Note sur les noms des fichiers

- L'adresse (raccourcie) http://home.etu.unige.ch/~votrelogin pointe sur le fichier index.html ou index.htm de votre dossier H:\PUBLIC.WWW\
- Cette URL devrait aussi pouvoir pointer sur le fichier index.xhtml si les deux autres n'existent pas/plus, mais ce n'est pas (encore) le cas
- Que faire si vous voulez que index.xhtml soit votre page par défaut, consultable comme http://home.etu.unige.ch/~votrelogin ?
  - solution : renommer votre page index.xhtml en index.html
    - pour qu'il n'y ait pas de problème d'affichage, enlever aussi la déclaration XML du début du fichier
    - tester toujours avec d'autres navigateurs, depuis d'autres postes
  - si vous ne voulez pas que index.xhtml soit votre page par défaut, il n'y a rien à faire, il faudra seulement indiquer l'URL complète

19