



Date : 21/07/2006

Comment la localisation gêne les portails internationaux : jeux de caractères et accès international

Pierre Clavel

Bibliothèque nationale suisse

Berne, Suisse

Meeting:	77 UNIMARC
Simultaneous Interpretation:	Yes
<p>1. WORLD LIBRARY AND INFORMATION CONGRESS: 72ND IFLA GENERAL CONFERENCE AND COUNCIL 20-24 August 2006, Seoul, Korea http://www.ifla.org/IV/ifla72/index.htm</p>	

Résumé:

L'accès à des ressources électroniques disponibles dans un autre pays que le sien peut poser des difficultés techniques. Cet article montre quels obstacles perturbent l'accès au niveau des jeux de caractères. La plupart des logiciels et certains composants matériels, tels les claviers, sont localisés, c'est-à-dire déclinés en versions adaptées à une langue et à son usage dans une région déterminée. Les usagers sont maintenant habitués à cela et le considèrent normal. Bien qu'Unicode facilite grandement l'interopérabilité, les portails internationaux rencontrent des difficultés dans ce domaine à cause de différences significatives entre langues et/ou pays dans le traitement des données, qui continuent à gêner la recherche, l'affichage et le tri de données.

Cette étude a été menée dans le cadre du projet TEL-ME-MOR, avec le soutien de la Commission européenne.

1. Introduction

Plusieurs facteurs ont contribué à l'accroissement du nombre de portails, plateformes accédant simultanément à plusieurs bases de données. La répartition géographique et linguistique de leurs partenaires a également augmenté. Un facteur de cette croissance, à côté des performances améliorées des réseaux et ordinateurs, est certainement la diminution du nombre de jeux de caractères employés, exigeant ainsi moins de conversions pour permettre l'interopérabilité. Il reste cependant un facteur mineur et parfois ignoré, le plus important étant la plus grande disponibilité de documents en ligne permettant un accès de partout dans le monde. On pourrait se demander si les différences de langue ne seraient pas l'obstacle annulant les possibilités offertes par les progrès évoqués ci-dessus. Bien que la barrière des langues subsiste, son impact est atténué par le fait la plupart des bibliothèques ont des documents dans d'autres langues que la langue locale et que de nombreux documents en ligne, visuels et sonores, ne sont pas liés à une langue. Ils doivent néanmoins être décrits et indexés, ce qui fait ressurgir les mêmes questions de jeux de caractères. Bien que cet article se concentre sur les portails de bibliothèques, The European Library

(www.theeuropeanlibrary.org) en particulier, les questions discutées ici concernent également d'autres types de portails et de moteurs de recherche.

Le but de The European Library est d'offrir l'accès à plusieurs ressources par la combinaison d'un index centralisé et de recherches distribuées. Ces ressources sont des documents numériques dans divers formats et des notices bibliographiques provenant actuellement de 16 bibliothèques nationales européennes. Le but final est d'intégrer toutes les bibliothèques nationales des 46 pays membres du Conseil de l'Europe (page d'entrée en français à l'adresse www.coe.int/DefaultFR.asp, disponible en 32 autres langues). Ceci n'impliquera pas moins de 5 systèmes d'écriture différents: arménien, cyrillique, géorgien, grec et latin. Cet article n'aborde les questions de jeux de caractères que dans le cas de l'alphabet latin.

Le but du projet TEL-ME-MOR (The European Library: Modular Extensions for Mediating Online Resources, www.telmemor.net) est d'aider les bibliothèques nationales des Nouveaux Etats Membres de l'Union européenne à devenir membres à part entière de The European Library. L'une en était déjà membre, (Slovénie), trois l'avaient rejointe au moment d'écrire ces lignes (Estonie, Lettonie, Slovaquie) et les six autres devraient y entrer d'ici au début de 2007. Six autres partenaires, dont la Bibliothèque nationale suisse, contribuent à des tâches spécifiques du projet pour lesquelles ils disposent de compétences techniques ou administratives particulières.

2. Notions sur les jeux de caractères et les problèmes qu'ils peuvent poser

Les ordinateurs mémorisent et traitent l'information par une succession de *bits*, éléments de mémoire pouvant n'avoir que deux valeurs, 0 et 1. Pour agrandir le nombre de valeurs possibles, les bits sont combinés en groupes où le nombre de valeurs possibles est égal à 2 à la puissance du nombre de bits, puisque chaque bit supplémentaire double le nombre de valeurs possibles. Exemples:

2 bits: 00, 01, 10, 11 = 4 valeurs = 2^2

3 bits: 000, 001, 010, 011, 100, 101, 110, 111 = 8 valeurs = 2^3

etc.

Un groupe de bits peut ainsi représenter un caractère. La succession des valeurs possibles associées aux caractères en fait un jeu de caractères.

Chaque constructeur des premiers ordinateurs avait défini son propre jeu de caractères, mais le besoin d'une normalisation s'est fait sentir à la fin des années 1950 pour permettre l'échange de données.

Le premier jeu de caractères normalisé, ASCII (American Standard Code for Information Interchange), avait 7 bits, donc 128 valeurs possibles, et permettait de stocker correctement des textes en anglais. À ce stade, les langues employant un autre système d'écriture ou l'alphabet latin avec des accents sur certaines lettres n'avaient simplement aucun moyen d'être traitées correctement. ASCII fut rendu international en 1967 comme norme ISO 646, avec la définition de 10 codes de caractères réservés pour des variantes nationales.

La transmission de données devint progressivement suffisamment fiable pour que l'on pût employer le huitième bit pour des codes supplémentaires plutôt que comme bit de parité, ouvrant la voie à l'usage d'ordinateurs dans n'importe quelle langue disposant d'un système d'écriture alphabétique.

On développa un grand nombre de nouvelles normes à 8 bits pour enrichir la gamme des caractères disponibles et répondre aux besoins de divers langues et groupes d'utilisateurs. Environ 180 normes incluent l'ASCII et sont compatibles entre elles pour la première moitié des jeux de caractères mais pas pour la seconde, comme le montre l'exemple ci-dessous.

Une personne travaillant sur un PC configuré pour l'Europe de l'Ouest écrit la phrase suivante dans un simple fichier texte

'Le théâtre sera réparé après le mois d'août, là où même le plâtre paraît neuf.'

Si ce fichier est ensuite transmis et ouvert sur un Macintosh ou un très vieux PC également configuré pour l'Europe de l'Ouest mais employant MS-DOS, ou encore sur un PC configuré pour l'Europe de l'Est, cette phrase se lira, respectivement:

'Le thÈ,tre sera rÈparÈ aprÈs le mois d'ao°t, l‡ o~ mlme le pl,tre paraÓt neuf.'

'Le thŒtre sera rŒparŒ aprŒs le mois d'ao√t, l‡ o· mΩme le plŒtre paraet neuf.'

'Le théâtre sera réparé aprĉs le mois d'août, lġ où mĉme le plâtre paraġt neuf.'

Quelqu'un connaissant la langue employée a peut-être de bonnes chances de rétablir les lettres correctes en les devinant. Cependant, en plus de pouvoir considérer qu'un tel affichage n'est pas satisfaisant, on voit que chercher dans un catalogue avec un jeu de caractères incorrect ne permettra pas d'obtenir de bons résultats.

3. Première étape vers une solution

Les logiciels permettent maintenant de charger et d'employer plusieurs jeux de caractères à 8 bits simultanément mais cela exige de signaler quelque part le jeu de caractères employé pour tout segment de texte, à moins qu'il ne s'agisse du jeu de caractères par défaut. Le jeu de caractères par défaut peut cependant être différent d'un ordinateur à l'autre, ce qui perturbe leur interopérabilité car annoncer son jeu de caractères par défaut n'est pas présent dans tous les protocoles de communication. C'est pourquoi l'industrie et les organes de normalisation recherchent une solution amenant à un jeu de caractères unique et complet. Après des premiers travaux effectués séparément (et même conflictuellement) Unicode (www.unicode.org, l'initiative de l'industrie) et ISO 10646 (l'initiative des organes de normalisation) se mirent d'accord sur des normes où au moins les codes des caractères sont identiques. Elles diffèrent seulement au niveau d'informations supplémentaires concernant les caractères, Unicode leur attribuant plus de propriétés qu'ISO.

Unicode est un jeu de caractères à 16 et 32 bits destiné à contenir dans un jeu unique tous les caractères nécessaires. 16 bits permettent de définir 65'536 caractères, bien assez pour incorporer tous les alphabets et les syllabaires. Ce nombre a été étendu à 1'114'112 en définissant une section virtuelle à 32 bits destinée aux idéogrammes. Comme les jeux de caractères à 8 bits contenant ASCII, Unicode inclut Latin-1 (ISO 8859-1), le rendant compatible avec le jeu à 8 bits de l'Europe de l'Ouest sans autre conversion que l'insertion d'octets vides.

Unicode et ISO 10646 décrivent des caractères, pas des glyphes. L'exemple suivant le montre mieux qu'une explication:

Caractère	Code	Nom
В	U+0412	Lettre majuscule cyrillique ve
Β	U+0392	Lettre majuscule grecque beta
B	U+0042	Lettre majuscule latine b

En cyrillique, grec et latin, les lettres majuscules ve, beta et, respectivement, b ont exactement la même apparence. Elles emploient le même *glyphe*. Elles restent cependant des *caractères* différents, avec des codes différents, puisqu'elles sont employées dans des systèmes d'écriture différents, ce qui se remarque lorsqu'on voit leurs minuscules: в, β et b respectivement.

Dans les jeux de caractères à 8 bits, la relation entre le code d'un caractère et la valeur numérique effectivement traitée et mémorisée par les ordinateurs est simple : c'est la même dans (presque) tous les cas parce que les octets restent une unité de base des microprocesseurs. Unicode fait par contre une distinction entre le code d'un caractère en tant qu'abstraction et comment ce dernier est exprimé en bits, ce qu'on appelle son *modèle de codage* ou le mécanisme de sérialisation des caractères. Le choix d'un modèle de codage dépend de ce que l'on doit faire avec les caractères, par exemple les garder en mémoire ou leur faire subir un certain traitement, et les contraintes spécifiques à ces actions. Il y a 3 principaux modèles de codage, nommés UTF-32, UTF-16 et UTF-8.

Dans UTF-32, chaque caractère est encodé en 32 bits, ou 4 octets. Son emploi n'est réellement pertinent qu'avec des idéogrammes. Dans UTF-16, chaque caractère est encodé en 16 bits, ou 2 octets, sauf les caractères dont le code est supérieur à 65'535 (hexadécimal FFFF), où 4 octets sont nécessaires. Cet encodage est préféré dans des applications où la position d'un caractère au sein d'une chaîne doit pouvoir être prédite indépendamment des caractères qui le précèdent. Dans UTF-8, chaque caractère est encodé 1 à 4 octets selon un algorithme spécial. Les caractères ASCII emploient 1 octet, tous les autres caractères européens (y compris les systèmes d'écriture non latins) et du Proche-Orient 2 octets et les systèmes d'écriture asiatiques ainsi que certains caractères spéciaux comme les symboles mathématiques 3 ou 4 octets. Cet encodage est préféré lorsque l'espace mémoire ou la vitesse de transmission importe et est de ce fait répandu sur la toile.

La norme Unicode définit également certaines caractéristiques techniques des caractères. Chaque caractère a :

- Un *nom* normalisé.
- Une *catégorie générale*, telle que lettre, nombre, signe de ponctuation etc.
- Une *classe combinatoire canonique* : des signes diacritiques ou similaires peuvent être ajoutés à des lettres dans de nombreuses langues pour marquer une intonation ou modifier le son de la lettre. Ces signes sont codés individuellement dans Unicode et ce paramètre définit notamment où ils sont placés par rapport à leurs lettres porteuses.

de la page 3 donnerait déjà un on indice. Ensuite, pour faciliter la saisie, un clavier virtuel peut être ajouté à la page où l'on saisit sa recherche, de manière que les usagers puissent saisir les caractères absents de leur clavier physique.

Une autre aide pourrait venir de notices d'autorité enrichies. Si les plus importants noms et termes contenant un caractère spécial indexé individuellement dans son propre profil local mais pas dans les autres ont des renvois à partir d'une forme non localisée, tous les usagers auront de meilleures chances d'atteindre les notices pertinentes, c'est-à-dire même les usagers ignorant les particularités des profils locaux.

Une troisième solution au problème des règles d'indexation différentes pourrait être un doublement des index dans les systèmes pour bibliothèques, qui auraient alors une série d'index localisés et une série d'index non localisés, c'est-à-dire où tous les signes diacritiques sont ignorés et tous les caractères spéciaux convertis en la lettre ordinaire la plus proche. Un port Z39.50 séparé et destiné à l'accès international et aux portails pointerait alors vers les index non localisés. On remarquera que cette solution échange le « silence » contre le « bruit », puisque certains noms ou mots différenciés dans certains profils locaux seraient indexés de la même manière. Pour cette raison, l'accès aux index non localisés devrait rester une option, permettant aux usagers de choisir l'accès le plus approprié à leurs connaissances linguistiques et aux caractéristiques de leur recherche.

Comme ceci n'est techniquement pas réalisable actuellement, The European Library appliquera la première solution, documentation des caractères spéciaux et clavier virtuel. La faiblesse de cette solution est que les usagers lisent rarement les instructions et que cela leur demande un niveau de connaissances qu'ils n'ont pas nécessairement. Heureusement, plusieurs partenaires appliquent la seconde solution et complètent une partie de leurs notices d'autorité par des formes non localisées.

La résolution des divergences des profils locaux sur l'ordre alphabétique est possible en faisant faire le tri par le portail, sur lequel il n'est pas difficile d'installer des modules de tri pour tous les profils locaux européens. On peut alors laisser l'utilisateur choisir le profil local qu'il désire voir appliquer à ses résultats ou, plus simplement, le déduire de la langue d'interface sélectionnée. Il serait nécessaire de retrier les ensembles de notices envoyés par certains systèmes, afin d'appliquer le profil local attendu par l'utilisateur.

La localisation est une bonne chose : elle rend les ordinateurs plus conviviaux et tient compte de la diversité culturelle. Il est cependant important de ne pas compromettre la coopération internationale dans l'accès aux ressources numériques. Passer au système uniforme des débuts à une série de profils locaux adaptés aux diverses communautés d'utilisateurs a en effet entraîné quelques nouveaux problèmes. Les portails internationaux ont un défi à relever dans ce domaine et, heureusement, quelques moyens pour le faire. Ceux-ci comprennent aussi bien des réglages techniques que la sensibilisation des usagers de cette diversité culturelle, en toute concordance avec les buts de l'Europe.