

Demain, encore plus de tfinaghés sur Internet

Patrick Andries

Conseils Hapax, Québec, Canada
Membre du consortium Unicode
patrick@hapax.qc.ca

Résumé. Lors de cette communication, nous nous pencherons sur des aspects de l'Internet où les tfinaghés sont actuellement peu présents : les adresses de domaine internet, les adresses de courriel, les identifiants XML et les polices téléchargeables par les navigateurs.

In this paper, we will review areas in Internet where tfinagh characters are currently underrepresented: Internet domain names, email addresses, XML identifiers and dynamically downloadable browser fonts.

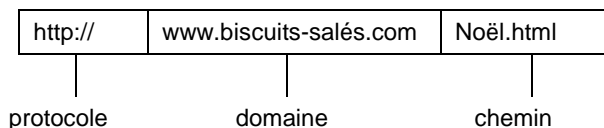
Mots-clés. Unicode, tfinaghe, ISO 10646, amazighe, informatique, réseaux, jeux de caractères, normalisation, polices, XML, noms de domaine, IDN, NDI, courriel, messagerie, touareg, arabe, navigateurs, internet.

1 Introduction

Il existe aujourd'hui encore plusieurs domaines d'Internet dans lesquelles il est difficile d'utiliser les tfinaghés, voire impossible. Nous allons nous pencher sur certains de ceux-ci et expliquer les évolutions les plus récentes dans ces domaines et les difficultés rencontrées qui limitent l'utilisation des tfinaghés dans ces applications.

2 Noms de domaine internationalisés (NDI)

Il est désormais permis d'avoir des adresses internationalisées du type <<http://www.biscuits-salés.com/Noël.html>>. Techniquement on nomme ce type d'adresse, des identificateurs de ressource internationalisée (IRI). Ces adresses se divisent en trois parties principales, la première avant le «://» indique le protocole à utiliser (ici http), ensuite vient le nom de domaine proprement dit puis, après un «/», le chemin de la ressource sur le serveur identifié par le nom de domaine.



Tous les domaines de tête (le *.ma*, le *.fr* ou *.com* final du nom de domaine) n'acceptent pas ces adresses internationalisées. Certains domaines de tête comme la Suisse (*.ch*) acceptent des accents français (et plus généralement les lettres accentuées de ses langues nationales) alors que la France n'accepte pas de tels accents dans les domaines qui se terminent par *.fr*. La politique d'attribution de ces noms est décidée par l'autorité d'enregistrement responsable d'un domaine de tête particulier. Au Canada, il s'agit d'une agence sans but lucratif (l'ACEI), elle est régie par les lois canadiennes. Les domaines de tête génériques (*.com*, *.org*, *.biz*) sont administrés aux États-Unis en vertu des lois américaines.

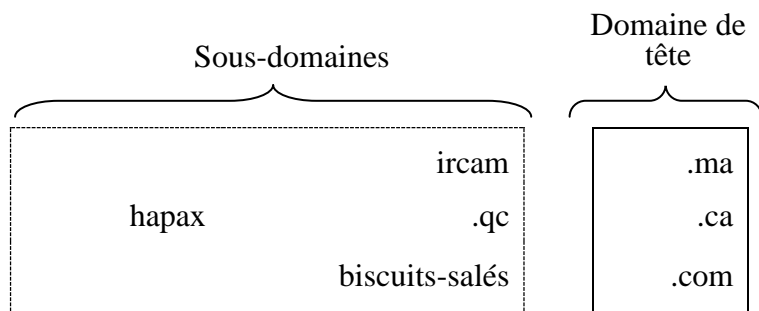


Figure 1. Parties d'un nom de domaine

Les navigateurs modernes transformeront de manière transparente les adresses de domaine comme *biscuits-salés* en des adresses de domaine compatibles avec l'infrastructure actuelle prévue pour des adresses ASCII. L'utilisateur ne se rendra habituellement pas compte des transformations effectuées en coulisse par le protocole appelé *punycode*. Dans notre cas, ce protocole transforme *biscuits-salés* en *xn-biscuits-sals-mhb* afin de pouvoir réutiliser toute l'infrastructure internet actuelle qui ne permet que l'utilisation d'ASCII. C'est sous ce nom étrange que sera enregistré le nom de domaine internationalisé auprès de l'autorité responsable pour l'attribution des sous-domaines pour un domaine de tête particulier (l'ACEI au Canada par exemple). Et

c'est sous cette forme que le nom sera diffusé aux différents serveurs de noms de domaine (les DNS) qui effectueront le passage entre le nom d'un domaine et son adresse IP qui n'est composée que de chiffres. Lors d'une demande de chargement d'une page HTML, seuls les navigateurs verront donc le nom de domaine internationalisé, tout le reste de l'infrastructure ne verra que de l'ASCII, sous la forme *punycode* si le nom de domaine contenait des caractères non ASCII. C'est ce qu'illustre la figure 2.

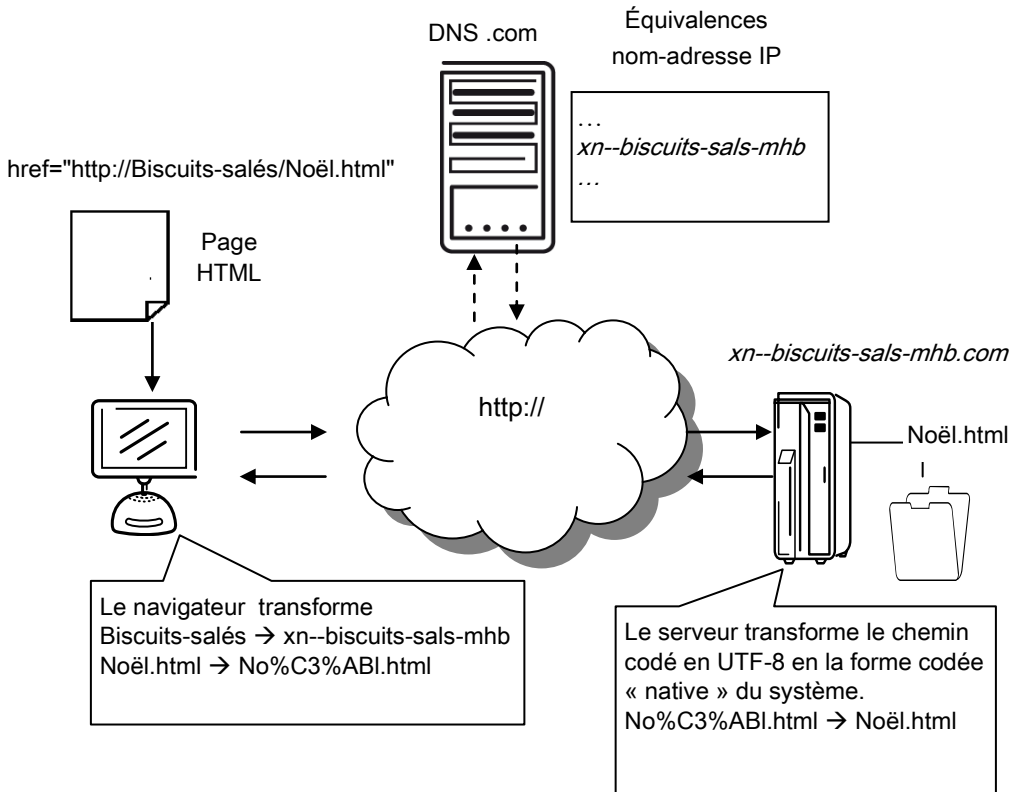


Figure 2. Transformation des IRI

Pour le codage du chemin de la ressource sur le serveur (à savoir Noël.html), le mieux est d'adopter le RFC 3987 qui préconise l'utilisation d'UTF-8 comme codage des chemins avant leur transformation en notation %hh et leur transmission sur le réseau : Noël.html devient donc No%C3%ABl.html.

La prise en charge des NDI peut augmenter le risque de parodie ou d'hameçonnage (« phishing ») par l'entremise de caractères similaires à ceux auxquels s'attend l'utilisateur, mais dans un autre système d'écriture. Une adresse ainsi contrefaite pourrait, par exemple, utiliser un « o » cyrillique (U+043E) à la place d'un « o » latin (U+006F) ! (Ces deux caractères ont le plus souvent une apparence identique.) La stratégie la plus souvent employée pour détecter ces adresses contrefaites consiste pour les navigateurs à ne pas permettre l'emploi dans une même partie du nom de domaine délimitée par un « . » de caractères provenant d'écritures différentes (pas de caractères latins et cyrilliques pour reprendre notre exemple dans un même segment). Les autorités d'enregistrement de noms mettent en œuvre des règles similaires pour éviter tout risque d'usurpation : elles sont souvent plus draconiennes encore que les navigateurs. C'est ainsi que l'autorité suisse d'enregistrement n'acceptera que les caractères latins utilisés par ses langues nationales au sein des noms de domaine qu'elle octroie. Toutefois, comme on l'a vu, ces noms de domaine pourront comprendre des accents et la cédille.

Malheureusement, en 2008, seul un sous-ensemble des caractères d'Unicode peut être utilisé dans les noms de domaine internationalisés. Ils doivent tous faire partie du répertoire d'Unicode 3.2, or ce répertoire ne comprend aucun caractère tiffinaghe puisque cette écriture n'a été incluse que dans la version 4.1 d'Unicode.

Toutefois, une nouvelle version de la norme qui régit les caractères permis dans les NDI est en préparation (IDNA2008) elle devrait permettre d'utiliser tous les caractères tiffinaghes d'Unicode.

3 Adresses de courriel internationalisées

Il existe un domaine voisin des noms de domaine internationalisés (NDI) où, pour l'instant, la prise en charge des caractères Unicode est incomplète. Il s'agit de celui des adresses de courriel internationalisées.

C'est ainsi que s'il est possible – pour autant que vos programmes prennent en charge les NDI – d'avoir des adresses de courriel de ce type :

pandries@hapax.qc.ca

hsu.wang@环球电邮.info

Il n'est pas possible à l'heure actuelle d'avoir des adresses où les caractères à la gauche de l'arrobe (@) sont des caractères accentués ou

appartenant à une écriture non latine comme dans les exemples ci-dessous :

françois.côté@biscuits-salés.com

测试@环球电邮.info

En effet, les adresses de courriel internationalisées ne sont pas compatibles avec le standard SMTP de base. Ainsi, la commande RCPT TO qui sert à préciser le destinataire ne permet que des caractères ASCII. Il faut donc étendre SMTP pour permettre l'utilisation de UTF-8 dans les commandes MAIL FROM et RCPT TO. Cette extension est l'objet du standard expérimental RFC 5336, elle se signale dans le protocole SMTP par l'envoi de l'option UTF8SMTP, dont l'usage permet de vérifier que le serveur de transmission de courriels avec lequel correspond votre logiciel comprend bien la nouvelle norme. Si ce n'est pas le cas, l'échange s'effectuera à l'aide d'adresses en ASCII.

Mais voyons à quoi ressemble un échange entre un client (votre logiciel de messagerie) et un serveur messagerie. Le client (C) appelle le serveur (S) dans le dialogue ci-dessous :

```
S: 220 amy.cooptel.qc.ca ESMTP
C: EHLO courriel.orange.be
S: 250-amy.cooptel.qc.ca
  250-ENHANCEDSTATUSCODES
  250-8BITMIME
  250-UTF8SMTP
C: MAIL FROM:<françois.côté@biscuits-salés.com> ALT-
  ADDRESS=fcote@biscuits-au-sel.com
S: 250 ok
C: RCPT TO: <测试@环球电邮.info>
S: 250 ok
```

Dans la première ligne, le serveur s'annonce et donne son adresse (amy.cooptel.qc.ca). Ensuite, à la ligne suivante, le client salue le serveur (EHLO) et fournit son nom de domaine (courriel.orange.be). Plus loin, le client précise qu'un message de François Côté doit être envoyé à 测试@环球电邮.info.

Notez que les deux adresses comprennent des caractères non ASCII, à gauche et à droite de l'arobres (@). Le MAIL FROM : contient deux adresses de l'expéditeur : une internationalisée et une autre ASCII de repli qui se conforme au protocole SMTP de base. L'adresse du destinataire peut également comprendre un paramètre ALT-ADDRESS

Il existe plusieurs remèdes possibles. Une solution consiste à prévenir les lecteurs des pages en tiffinage qu'ils doivent installer une ou plusieurs polices en fournissant un lien permettant de télécharger ces polices. Il existe cependant une autre solution : les polices dynamiquement téléchargeables. Cette solution consiste à envoyer les glyphes nécessaires à l'affichage d'une page HTML avec la page en question.

5.1 Police téléchargeable avec Internet Explorer

Les polices téléchargeables ne sont pas une nouveauté : Internet Explorer 4 le permettait déjà en 1997. Comme nous le verrons, ci-dessous, ce qui a changé c'est le fait que d'autres navigateurs Internet offrent désormais cette fonctionnalité. Mais revenons à Internet Explorer.

Microsoft offre un programme qui se nomme WEFT et qui permet d'analyser, d'une part, une page HTML et, d'autre part, une police TrueType incorporable. C'est-à-dire une police dont l'auteur permet qu'elle accompagne un document.

Sur la base de ces deux sources, WEFT produit une police téléchargeable (dont le suffixe est .eot) et un extrait de feuille de style CSS qui fait référence à cette police .eot. L'illustration ci-dessous décrit graphiquement ce processus.

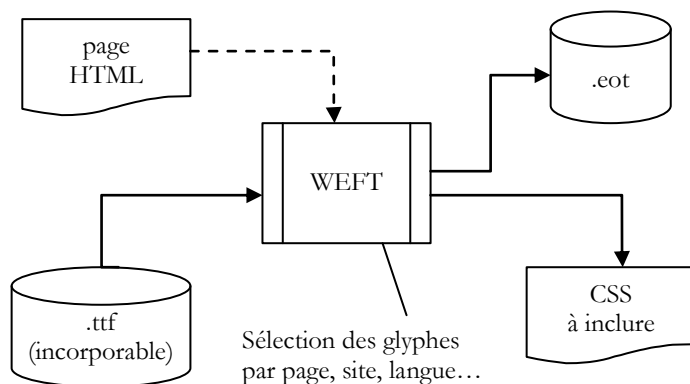


Figure 4. WEFT de Microsoft

Le fichier .eot ne doit pas nécessairement comprendre tous les glyphes de la police TrueType fournie en entrée, c'est le plus souvent inutile et inefficace. WEFT permet ainsi de sélectionner les seuls glyphes nécessaires à l'affichage d'une page HTML ou d'un ensemble de pages.

La police .eot est chiffrée et idéalement ne comprend donc que le sous-ensemble de glyphes nécessaire. Pour des raisons de sécurité, le fichier .eot produit est lié à une liste de serveurs explicites et ne peut être utilisés qu'à partir de ceux-ci. Le plus gros désavantage lié à ce mécanisme est cependant le fait que ces fichiers ne fonctionnent que sur Windows. L'extrait de code ci-dessous représente le code CSS que WEFT produit et qui doit être inclus dans les fichiers HTML fournis en entrée pour qu'ils invoquent la police .eot stockée.

```
<style type="text/css">

    @font-face {
        font-family: "Ajoure-eot";
        font-style: normal;
        font-weight: normal;
        src:
        url(http://hapax.qc.ca/polices/AJOURE0.eot);
    }

    body { font-family: "Ajoure-eot"; }

</style>
```

Une fois la page HTML de la figure 3 modifiée pour inclure ce code CSS, le texte tiffnaghe apparaîtra correctement dans Internet Explorer pour peu que la police .eot mentionnée dans la partie `src :url()` du `@font-face` soit bien présentée à l'adresse mentionnée.

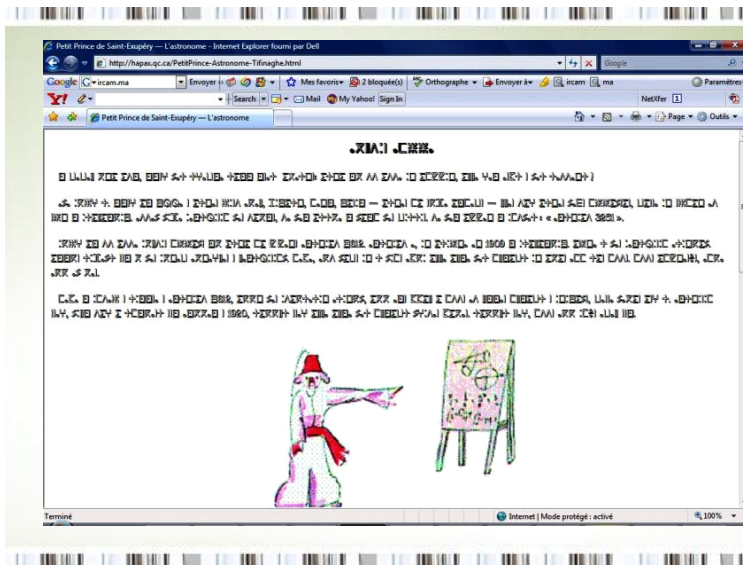


Figure 5. Affichage de la police tiffnaghe avec MS Explorer

5.2 Les autres navigateurs

Malheureusement, les autres navigateurs internet comme Firefox, Opera, Safari ou Chrome ne comprennent pas les fichiers `.eot` dont le format est propriétaire.

Toutefois, `@font-face` permet également de préciser le chargement d'une police TrueType (`.ttf`) ou OpenType (`.otf`). Les dernières versions de Firefox 3.1 (Windows) et Safari 3.1 (sur Mac et Windows) prennent en charge `@font-face`.

Il est dès lors possible d'écrire une feuille de style qui fonctionne sur plusieurs navigateurs grâce à du code comme celui que l'on retrouve ci-dessous :

```
<style type="text/css" >
  @font-face {
    font-family: "Ajoure-ttf";
    font-style: normal;
    font-weight: normal;
    src: url(http://hapax.qc.ca/polices/ajoure.ttf);
  }
  @font-face {
    font-family: "Ajoure-eot";
    font-style: normal;
    font-weight: normal;
    src: url(http://hapax.qc.ca/polices/AJOURS0.eot);
  }

  body {font-family: "Ajoure-ttf","Ajoure-eot";}
  h1 {font-size: 16pt; text-align:center; space-before:
12pt; space-after: 12pt;}
  p {text-indent:9pt;}

</style>
```

Dans un premier temps, on trouve deux `@font-face`. L'un pour les navigateurs internet qui prennent en charge le chargement des polices TrueType, l'autre pour MS Explorer qui met en œuvre les polices « `.eot` ».

Ensuite, il suffit de préciser en regard du sélecteur du corps complet du document (`body {...}`), les deux polices téléchargeables équivalentes, l'une sous le format `.ttf` et l'autre sous le format `.eot`. Le navigateur choisira la première police disponible où il trouvera les glyphes

nécessaires. Dans le cas de Safari ce sera dans la police .ttf, pour ce qui est de Microsoft Explorer ce sera dans la police .eot.

Grâce à ce code CSS et après avoir hébergés les polices correspondantes à l'adresse référencée par leur `src:url()` correspondant, on peut désormais afficher la même page HTML tifnaghe dans Safari. Et cela, sans que l'utilisateur n'ait besoin d'installer au préalable la bonne police tifnaghe.

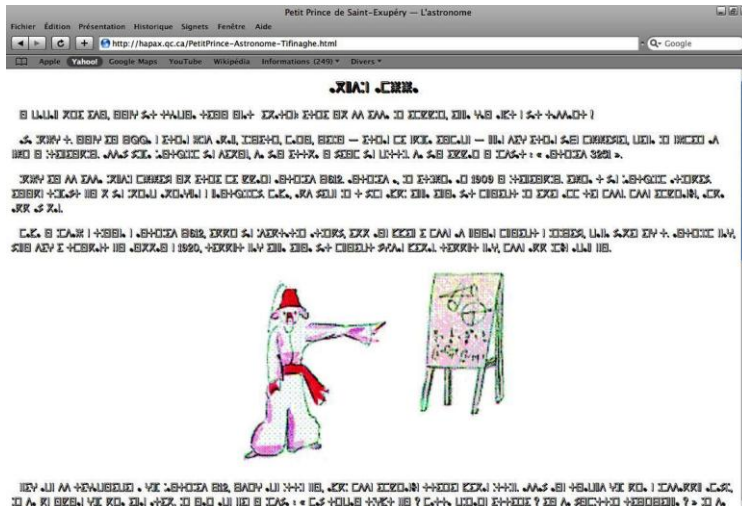


Figure 6. Affichage de la police tifnaghe avec Safari

5.3 À plus long terme...

Pour l'instant, les polices TrueType sont envoyées d'un bloc alors que les polices .eot peuvent n'envoyer que les glyphes nécessaires à l'affichage de la page HTML correspondante, allégeant d'autant la transmission de la police incorporée.

Le W3C considère actuellement un mécanisme ouvert de même type que .eot, il devrait être valable pour toutes les plateformes, il sera associé à des pages d'un site, pourra produire des sous-ensembles de glyphes sélectionnés et compresser ces polices téléchargeables.

6 Conclusion

Depuis 2005 et la normalisation du tifnaghe dans Unicode et l'ISO/CEI 10646, beaucoup de choses ont changé. Il est de plus en plus facile d'utiliser des tifnaghes sur internet. Il reste malheureusement quelques obstacles à leur emploi généralisé comme nous l'avons vu dans

le cas des noms de domaine internationalisés, en XML, dans les adresses courriels et dans l'affichage de pages sur Internet quand une police tifinaghe manque, ce qui est encore souvent le cas.

Toutefois, dans chacun de ces domaines, les choses avancent et des solutions techniques se dessinent pour que l'utilisation des tifinaghes dans tous ces domaines se fasse désormais sans entraves et cela principalement grâce à l'inclusion de cette écriture dans Unicode.

7 Remerciements

Nous tenons à vivement remercier l'IRCAM et plus particulièrement le directeur du CEISIC, Youssef Aït Ouguengay, pour leur accueil chaleureux et l'organisation du colloque international à Rabat au cours duquel cette communication a été présentée. Nous voulons également ici rendre hommage au prédécesseur de M. Aït Ouguengay, le professeur Lahbib Zenkour, sans lequel la normalisation informatique du tifinaghe n'aurait été ni aussi rapide ni aussi complète.

8 Bibliographie

Andries, P. (2008). *Unicode 5.0 en pratique*, Dunod éditions, Paris.

Bortzmeyer, S. (2008), *RFC 5336: SMTP extension for internationalized email address*, disponible à <<http://www.bortzmeyer.org/5336.html>>.

Yao, J. et Mao W. (2008), *RFC 5336, SMTP Extension for Internationalized Email Addresses*, disponible à <<http://www.ietf.org/rfc/rfc5336.txt>>