RFC 6532 : Internationalized Email Headers

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 18 février 2012

Date de publication du RFC : Février 2012

https://www.bortzmeyer.org/6532.html

Dans l'ensemble <https://www.bortzmeyer.org/courrier-entierement-internationalise.html> des RFC qui décrivent les adresses de courrier électronique internationalisées (c'est-à-dire pouvant utiliser tout le répertoire Unicode), celui-ci se consacre au nouveau format des en-têtes, complétant celui normalisé dans le RFC 5322 [1].

Désormais, on peut avoir dans un message un en-tête comme :

```
From: Stéphane Bortzmeyer <stéphane@sources.org>
```

Oui, avec du vrai UTF-8, y compris dans l'adresse proprement dite, pas seulement dans les commentaires, et encore, à condition de les surencoder selon le RFC 2047 comme c'était le cas avant (méthode inélégante et inefficace). Cette possibilité d'utiliser directement UTF-8 avait été créée par le RFC 5335, qui n'avait que le statut « Expérimental ». Désormais, ce mécanisme est complètement normalisé, en partie parce que de plus en plus de logiciels savent traiter directement l'UTF-8.

L'un des éditeurs du RFC travaille à TWNIC <http://www.twnic.tw/>, le registre chinois de .tw. Les Chinois ont, fort logiquement, été particulièrement actifs dans tout le processus EAI ("*Email Addresses Internationalization*"). Celui-ci visait à terminer l'internationalisation du courrier électronique, qui avait passé une étape décisive en novembre 1996, avec la sortie du RFC 2045 sur MIME. Mais cette norme MIME n'internationalisait que le **corps** des messages. EAI permet d'internationaliser également les **en-têtes** comme `From:`, `Subject:` ou `Received:` (mais pas, par exemple, `Date:`, qui restera en ASCII pur ; `Date:` est censé être analysé par le MUA et présenté ensuite à l'utilisateur dans sa langue).

1. Pour voir le RFC de numéro NNN, https://www.ietf.org/rfc/rfcNNN.txt, par exemple https://www.ietf.org/rfc/rfc5322.txt

EAI comprend plusieurs RFC, par exemple le RFC 6530 définit le cadre général, le RFC 6531, l'utilisation des adresses Unicode dans SMTP (avec l'extension `SMTPUTF8`), les RFC ou futurs RFC sur POP (RFC 5721) ou IMAP (RFC 5738) et notre RFC 6532, qui se concentre sur le format des messages. Il suppose que le transport sera « "*8-bits clean*" », c'est-à-dire laissera passer les octets qui codent les données UTF-8 sans les modifier (le cas des vieux systèmes qui ne gèrent pas proprement les caractères non-ASCII est délibérement laissé de côté).

Concrètement, que change donc ce RFC ? La section 3 détaille ce qui est modifié, par extension de la grammaire du RFC 5322) :

— La grammaire ABNF qui décrit les en-têtes est changée (sections 3.1 et 3.2) pour accepter des caractères UTF-8, de préférence normalisés en NFC (cf. RFC 5198). Déjà très complexe, cette grammaire devient ainsi réellement difficile.
— L'en-tête `Message-ID:` est traité à part (section 3.3). Les caractères UTF-8 y sont autorisés mais déconseillés (le RFC rappelle que cet en-tête sert à former automatiquement d'autres en-têtes comme `In-Reply-To:`).
— La syntaxe des adresses (telles qu'on les trouve dans des en-têtes comme `From:` ou `To:`) est modifiée pour accepter UTF-8. Pour les sites qui utilisent une adresse électronique comme identificateur, comme le fait Amazon.com, ce sera un intéressant défi !
— Un nouveau type MIME est créé, `message/global`, pour permettre d'identifier un message EAI (section 3.7). Les messages de ce type ne doivent être transportés qu'avec des protocoles "*8-bits clean*" comme celui du RFC 6531.
— Les **noms** des en-têtes restent en ASCII seul, ce n'est que leur contenu qui peut être en UTF-8. Pour utiliser le vocabulaire du RFC 2277, les noms des en-têtes sont un élement de protocole, qui n'a pas besoin d'être exprimé dans la langue de l'utilisateur.

Parmi les conséquences pratiques de ce changement, il faut noter que les programmeurs qui écrivent des logiciels traitant le courrier doivent désormais gérer Unicode, même s'ils se contentent d'analyser les en-têtes... L'époque du courrier traité avec des scripts shell est bien passée.

Autre point important pour les programmeurs, les limites de taille des en-têtes (qui étaient de 998 caractères maximum et 78 recommandés) changent de sémantique (section 3.4) puisque, en UTF-8, un caractère ne fait plus forcément un octet. Les limites sont désormais de 998 **octets** et 78 **caractères** (cette seconde limite étant conçue pour tenir sur l'écran). Attention aussi à la section 4, qui rappelle le risque de débordement de tampons.

La section 4 liste également quelques conséquences pour la sécurité comme le fait qu'un mécanisme d'authentification devra être prêt à gérer plusieurs adresses pour la même personne (car, pendant un certain temps, plusieurs utilisateurs auront à la fois une adresse en Unicode et une en ASCII pur).

Les changements depuis le RFC 5335 sont résumés dans la section 6 du RFC 6530. Le principal est la suppression de la possibilité de **repli** ("*downgrading*") automatique vers des adresses en ASCII pur.