

RFC 5738 : IMAP Support for UTF-8

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 4 Mars 2010

Date de publication du RFC : Mars 2010

<http://www.bortzmeyer.org/5738.html>

Presque toutes les normes produites par l'IETF sont désormais complètement internationalisées. Il reste quelques trous par-ci, par-là, qui sont comblés lentement. Par exemple, notre RFC 5738¹ normalise le support d'Unicode par le protocole IMAP d'accès aux boîtes aux lettres.

Normalisé dans le RFC 3501, IMAP permet d'accéder à des boîtes aux lettres situées sur un serveur distant. Ces boîtes peuvent désormais avoir des noms en Unicode, les utilisateurs peuvent utiliser Unicode pour se nommer et les adresses gérées peuvent être en Unicode. L'encodage utilisé est UTF-8 (RFC 3629). Ce RFC 5738 fait donc partie de la série de RFC du groupe de travail EAI <<http://tools.ietf.org/wg/eai>> qui normalise un courrier électronique complètement international <<http://www.bortzmeyer.org/courrier-entierement-internationalise.html>>. Comme le changement est d'importance, ces RFC sont malheureusement encore étiquetés « expérimentaux ».

Tout commence par la possibilité d'indiquer le support d'UTF-8. Un serveur IMAP, à l'ouverture de la connexion, indique les extensions d'IMAP qu'il gère et notre RFC en crée une nouvelle, UTF8=ACCEPT (section 3). Par le biais de l'extension ENABLE (RFC 5161), le client peut à son tour indiquer qu'il va utiliser UTF-8. La section 3.1 détaille la représentation des chaînes de caractères UTF-8 sur le réseau.

Des commandes IMAP prennent désormais un paramètre UTF8. C'est le cas de SELECT et EXAMINE (section 3.2) qui permettent de manipuler les boîtes aux lettres. L'utilisation de ce paramètre par le client lors de la sélection d'une boîte change la sémantique de certaines commandes. Par exemple, toute commande qui indique la taille de la boîte aux lettres doit désormais compter en caractères Unicode et non plus en octets.

1. Pour voir le RFC de numéro NNN, <http://www.ietf.org/rfc/rfcNNN.txt>, par exemple <http://www.ietf.org/rfc/rfc5738.txt>

Les nouvelles capacités sont toutes décrites dans la section 10 et enregistrées dans le registre IANA <<http://www.iana.org/assignments/imap4-capabilities>>.

On peut désormais imaginer des boîtes aux lettres qui ne puissent être manipulées qu'en UTF-8. Dans ce cas, elles ne doivent **pas** être indiquées par défaut dans le résultat d'une commande `LIST` (sauf extensions de celle-ci, cf. RFC 5258), puisqu'elles ne pourraient pas forcément être sélectionnées ensuite. À noter qu'IMAP disposait depuis longtemps d'une astuce pour représenter les boîtes aux lettres dont les noms comportaient des caractères non-ASCII, la "*IMAP Mailbox International Naming Convention*" (RFC 3501, section 5.1.3). Elle n'a jamais donné satisfaction et la section 3.3 de notre RFC 5738 recommande de l'abandonner.

Il n'y a bien sûr pas que les boîtes, il y a aussi les noms d'utilisateurs qui peuvent être en Unicode (capacité `UTF8=USER`), et la section 5 spécifie ce point, en demandant que ces noms soient canonicalisés avec SASLprep (RFC 4013). Le RFC note (annexe A) que le serveur IMAP s'appuie souvent sur un système d'authentification externe (comme `/etc/passwd` sur Unix) et que, de toute façon, ce système n'est pas forcément UTF-8.

Aujourd'hui, rares sont les serveurs IMAP qui gèrent l'UTF-8. Mais, dans le futur, on peut espérer que l'internationalisation devienne la norme et la limitation à US-ASCII l'exception. Pour cet avenir radieux, la section 7 du RFC prévoit une capacité `UTF8=ONLY`. Si le serveur l'annonce, cela indique qu'il ne gère plus l'ASCII seul, que tout est en UTF-8 (un tel serveur, en 2010, n'aurait guère de clients...)

Outre les noms des boîtes et ceux des utilisateurs, cette norme IMAP UTF-8 permet à un serveur de stocker et de distribuer des messages dont les en-têtes sont en UTF-8, comme le prévoit le RFC 5335. Si son client ne gère pas UTF-8, le serveur doit alors (section 9) dégrader le messages en ASCII seul, comme expliqué dans le RFC 5504.

Chose relativement rare dans un RFC, l'annexe A contient une discussion détaillée des choix effectués et des raisons de ces choix. Par exemple, le fait que l'accès aux boîtes en UTF-8 se configure boîte par boîte et pas pour tout le serveur, a pour but de permettre aux serveurs existants de ne pas avoir à convertir toutes les boîtes stockées en UTF-8, lors de la mise à jour du logiciel.

Je ne connais pas encore d'implémentation en logiciel libre.